Weakly-Supervised Semantic Segmentation via Sub-category Exploration

Yu-Ting Chang[1] Qiaosong Wang[2] Wei-Chih Hung[1] Robinson Piramuthu[2]
Yi-Hsuan Tsai[3] Ming-Hsuan Yang[1,4]
[1]UC Merced [2]eBay Inc. [3]NEC Labs America [4]Google Research

Abstract

Existing weakly-supervised semantic segmentation methods using image-level annotations typically rely on initial responses to locate object regions. However, such response maps generated by the classification network usually focus on discriminative object parts, due to the fact that the network does not need the entire object for optimizing the objective function. To enforce the network to pay attention to other parts of an object, we propose a simple yet effective approach that introduces a self-supervised task by exploiting the sub-category information. Specifically, we perform clustering on image features to generate pseudo sub-categories labels within each annotated parent class, and construct a sub-category objective to assign the network to a more challenging task. By iteratively clustering image features, the training process does not limit itself to the most discriminative object parts, hence improving the quality of the response maps. We conduct extensive analysis to validate the proposed method and show that our approach performs favorably against the state-of-the-art approaches.

1. Introduction

The goal of semantic segmentation is to assign a semantic category to each pixel in the image. It has been one of the most important tasks in computer vision that enjoys a wide range of applications such as image editing and scene understanding. Recently, deep convolutional neural network (CNN) based methods [16, 5, 42] have been developed for semantic segmentation and achieved significant progress. However, such approaches rely on learning supervised models that require pixel-wise annotations, which take extensive effort and time. To reduce the effort in annotating pixel-wise ground truth labels, numerous weakly-supervised methods are proposed using various types of labels such as image-level [1, 22, 29, 32], video-level [6, 45, 35], bounding box [28, 8, 20], point-level [2], and scribble-based [26, 37] labels. In this work, we focus on using image-level labels which can be obtained effortlessly, yet a more challenging case under the weakly-supervised setting.

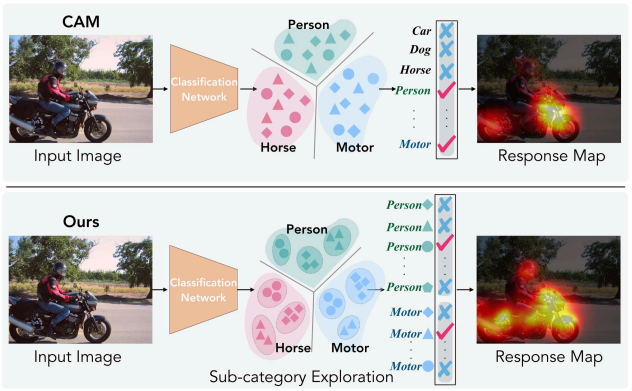

Figure 1: Existing weakly-supervised semantic segmentation methods based on image-level supervisions usually apply the class activation map (CAM) to obtain the response map as the initial prediction. However, this response map can only highlight the discriminative parts of the object (top). We propose a self-supervised task via sub-category exploration to enforce the classification network learn better response maps (bottom).

Existing algorithms mainly consist of three sequential steps to perform weakly-supervised training on the image-level label: 1) predict an initial category-wise response map to localize the object, 2) refine the initial response as the pseudo ground truth, and 3) train the segmentation network based on pseudo labels. Although promising results have been achieved by recent methods [1, 18, 38, 40], most of them focus on improving the second and the third steps. Therefore, these approaches may suffer from inaccurate predictions generated in the first step, i.e., initial response. Here, we aim to improve the performance of initial predictions which will benefit succeeding steps.

In order to predict the initial response map for each category, numerous approaches based on the class activation map (CAM) model [46] have been developed. Essentially, these methods train a classification network and use its learned weights in the classifier as the cues to compute

weighted sums of feature maps, which can be treated as the response map. However, such response maps may only focus on a portion of the object, instead of localizing the entire object (see top of Figure 1). One explanation is that the objective of the classifier does not need to "see" the entire object for optimizing the loss function. This impairs the classifier's ability to locate the objects.

At the core of our technique is to impose a more challenging task to the network for learning better representations, while not jeopardizing the original objective. To this end, we propose a simple yet effective method by introducing a self-supervised task that discovers sub-categories in an unsupervised manner, as illustrated at the bottom of Figure 1. Specifically, our task consists of two steps: 1) perform clustering on image features extracted from the classification network for each annotated parent class (e.g., 20 parent classes on the PASCAL VOC 2012 dataset [11]), and 2) use the clustering assignment for each image as the pseudo label to optimize the sub-category objective.

On one hand, the parent classifier establishes a feature space through supervised training as the guidance for unsupervised sub-category clustering. On the other hand, the sub-category objective provides additional gradients to enhance feature representations and leverage the sub-space of the original feature space to obtain better results. As such, the classification model takes a more challenging task and is not limited to the easier objective of learning only the parent classifier. Moreover, to ensure better convergence in practice, we iteratively alter the two steps of feature clustering and pseudo training the sub-category objective.

We conduct extensive experiments on the PASCAL VOC 2012 dataset [11] to demonstrate the effectiveness of our method, with regard to generating better initial response maps to localize objects. As a result, our approach leads to favorable performance for the final semantic segmentation results against state-of-the-art weakly-supervised approaches. Furthermore, we provide extensive ablation studies and analysis to validate the robustness of our method. Interestingly, we notice that the network is able to differentiate sub-categories with respect to their object size/type, context, and coexistence with other categories. The main contributions of this work are summarized as follows:

- We propose a simple yet effective method via a self-supervised task to enhance feature representations in the classification network. This improves the initial class activation maps for weakly-supervised semantic segmentation as well.
- We explore the idea of sub-category discovery via iteratively performing unsupervised clustering and pseudo training on the sub-category objective in a self-supervised fashion.
- We present extensive study and analysis to show the efficacy of the proposed method, which significantly improves the quality of initial response maps and leads to better semantic segmentation results.

2. Related Work

Within the context of this work, we discuss methods for weakly-supervised semantic segmentation (WSSS) using image-level labels, including approaches that focus on initial prediction and refinement for generating pseudo ground truths. In addition, algorithms that are relevant to unsupervised representation learning are discussed in this section.

Initial Prediction for WSSS. Initial cues are essential for segmentation task since it can provide reliable priors to generate segmentation maps. The class activation map [46] is a widely used technique for localizing the object. It can highlight class-specific regions that are served as the initial cues. However, since the CAM model is trained by a classification task, it tends to activate to the small discriminative part of the object, leading to incomplete initial masks.

Several methods have been developed to alleviate this problem. Numerous approaches [34, 39] deliberately hide or erase the region of an object, forcing models to seek more diverse parts. However, those methods either hide fixed-size patches randomly or require repetitive model training and response aggregation steps. A number of variants [44, 25] have been proposed to extend the initial response via an adversarial erasing strategy in an end-to-end training manner, yet such strategies may gradually expand their attention to non-object regions, leading to inaccurate attention maps. Recently, the SeeNet approach [17] applies self-erasing strategies to encourage networks to use both object and background cues, which prevent the attention from including more background regions. Instead of using the erasing scheme, the FickleNet method [24] introduces stochastic feature selection to obtain diverse combinations of locations on feature maps. By aggregating the localization maps, they acquire the initial cue that contains a larger region of the object.

Different from the methods that mitigate the problem by discovering complementary regions via iterative erasing steps or consolidating attention maps, our proposed approach aims at enforcing the network to learn harder on a more challenging task via self-supervised sub-category exploration, thereby enhancing feature representations and improving the response map.

Response Refinement for WSSS. Numerous approaches [1, 12, 13, 18, 22, 38, 40] are proposed to refine the initial cue via expanding the region of attention map. The SEC method [22] proposes a loss function that constrains both global weighted rank pooling and low-level boundary to

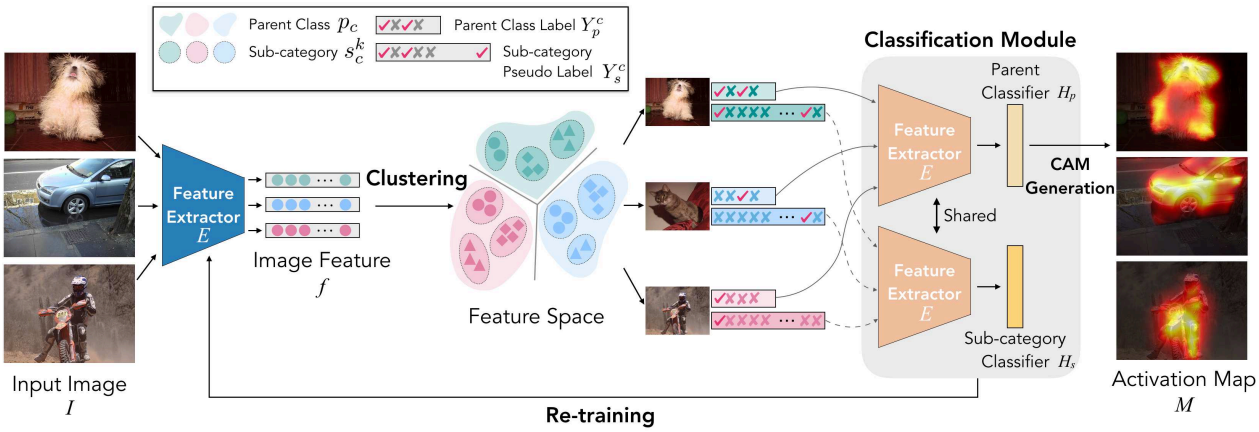

Figure 2: Proposed framework for generating the class activation map. Given input images I, we first feed them into a feature extractor E to obtain their features f. Then, we adopt unsupervised clustering on f and obtain sub-category pseudo labels Y_s for each image. Next, we train the classification network to jointly optimize the parent classifier H_p with ground truth labels Y_p for parent classes and the sub-category classifier H_s using the sub-category pseudo labels obtained in the clustering stage. By iteratively performing unsupervised clustering on image features and pseudo training the classification module, we use the jointly optimized classification network to produce the final activation map M.

expand the localization map. To improve the network training, the MCOF scheme [38] uses a bottom-up and top-down framework which alternatively expands object regions and optimize the segmentation network, while the MDC method [40] expands the seeds by employing multiple branches of convolutional layers with different dilation rates. Moreover, the DSRG approach [18] refines initial localization maps by applying a seeded region growing method during the training of the segmentation network. Other approaches are developed via affinity learning. For instance, the AffinityNet [1] considers pixel-wise affinity to propagate local responses to nearby areas, while [12, 13] explore cross-image relationships to obtain complementary information that can infer the predictions.

Nevertheless, initial seeds are still obtained from the CAM method. If these seeds only come from the discriminative parts of objects, it is difficult to expand regions into non-discriminative parts. Moreover, if the initial prediction produces wrong attention regions, applying the refinement step would cover even more inaccurate regions. In this paper, we focus on improving the initial prediction, which leads to more accurate object localization and benefits the refinement step.

Unsupervised Representation Learning. Unsupervised learning has been widely studied in the computer vision community. One advantage is to learn better representations of images and apply learned features on any specific domain or dataset where annotations are not always available. Self-supervised learning [9] utilizes a pretext task to replace the labels annotated by humans with "pseudo-labels" directly computed from the raw input data. A number of methods [27, 30, 31] are developed but require expert knowledge to carefully design a pretext task that may lead to good transferable features. To reduce the domain knowledge requirement, Coates and Ng [7] validate that feature-learning systems with K-means can be a scalable unsupervised learning module that can train a model of the unlabeled data for extracting meaningful features. Furthermore, a recent approach [3] employs a clustering framework to extract useful visual features by alternating between clustering the image descriptors and updating the weights of the CNN by predicting the cluster assignments, in order to learn deep representations specific to domains where annotations are scarce. In this work, we propose to learn a self-supervised method that explores the sub-category in the classification network, i.e., using unsupervised signal to enhance feature representations while improving initial response maps for weakly-supervised semantic segmentation.

3. Weakly-supervised Semantic Segmentation

In this section, we describe our framework for weakly-supervised semantic segmentation, including details of how we explore sub-categories to improve initial response maps and generate final semantic segmentation results.

3.1. Algorithm Overview

To obtain the initial response, we follow the common practice of training a classification network and utilize the CAM method [46] to obtain our baseline model. The CAM method typically only activates on discriminative object parts, which are not sufficient for the image classification

task. To address this issue, We propose to integrate a more challenging task into the objective: self-supervised sub-category discovery, in order to enforce the network to learn from more object parts.

Firstly, for each annotated parent class, we determine K sub-categories by applying K-means clustering on image features. With the clustering results, we then assign each image with a pseudo label, which is identified as the index of the sub-category. Finally, we construct a sub-category objective to jointly train the classification network. By iteratively updating the feature extractor, two classifiers, and sub-category pseudo labels, the enhanced features representations lead to better classification, and thereby gradually produce response maps that attain to more complete regions of the objects. The overall process is illustrated in Figure 2. Then, we use the method in [1] to expand response maps, which are used as pseudo ground truths to train the segmentation network. Also note that, our method focuses on the initial prediction, so it is not limited to certain region expansion or segmentation training methods.

Preliminaries: Initial Response via CAM. We adopt the CAM to generate the initial response using a typical classification network, whose architecture consists of convolutional layers as the feature extractor E, followed by global average pooling (GAP) and one fully-connected layer H_p as the output classifier. Given an input image I, the network is trained with image-level labels Y_p using a multi-label classification loss $\mathcal{L}_p$, following [46]. After training, the activation map M for each category c can be obtained via directly applying classifier H_p on the feature maps $f = E(I)$:

$$M^c(x, y) = \theta_p^{c\top} f(x, y), \tag{1}$$

where θ_p^c is the classifier weight for the category c, and $f(x, y)$ is the feature at pixel (x, y). The response map is further normalized by the maximum value in M^c.

3.2. Sub-category Exploration

The activation map for each image using (1) provides typically highlights only the discriminative object parts. However, from the perspective of a classifier, discovering the most discriminative part of the object is already sufficient for optimizing the loss function $\mathcal{L}_p$ in classification. As the learning objective is based on the classification scores, it is inevitable for the CAM model to generate incomplete attention maps. To address this issue, we integrate a self-supervised scheme to enhance feature representations f while improving the response maps via exploring the sub-category information, in which f appears to be an important cue to compute the activation map via (1).

Sub-Category Objective. To assign a more challenging problem to the classification model, we introduce a task to discover sub-categories in an unsupervised manner. For each parent class p_c, we define K sub-categories s_c^k, where $k = \{1, 2, , K\}$. For each image I with the parent label Y_p^c in $\{0, 1\}^c$, the corresponding sub-category label for the category c is denoted as $Y_s^{c,k}$ in $\{0, 1\}^k$. We also note that, if the label of one parent class does not exist (i.e., $Y_p^c = 0$), the labels of all sub-categories would be also 0, i.e., $Y_s^{c,k} = 0, k = \{1, 2, ..., K\}$. Our objective is to learn a sub-category classifier H_s parameterized with θ_s, while sharing the same feature extractor E with H_p. Similar to the parent classification loss $\mathcal{L}_p$, we adopt the standard multi-label classification loss $\mathcal{L}_s$ with a larger and fine-grained label space Y_s.

Sub-category Discovery. As there is no ground truth label for sub-category to directly optimize the above sub-category objective $\mathcal{L}_s$, we generate pseudo labels via unsupervised clustering. Specifically, we perform clustering for each parent class on image features extracted from the feature extractor E. The clustering objective for each class c can be written as:

$$\min_{D \in \mathbb{R}^{d \times k}} \frac{1}{N^c} \sum_{i=1}^{N^c} \min_{Y_s^c} ||f - TY_s^c||_2^2, \quad \text{s.t.}, Y_s^{c\top} 1_k = 1, \tag{2}$$

where T is a $D \times K$ centroid matrix, N^c is the number of images containing the class c, and $f = E(I) \in \mathbb{R}^D$ is the extracted feature. We use the clustering assignment Y_s^c for each image as the sub-category pseudo label to optimize $\mathcal{L}_s$.

Joint Training. After obtaining sub-category pseudo labels Y_s from the above clustering process, we jointly optimize the feature representations $f = E(I)$ and two classifiers, i.e., H_p and H_s:

$$\min_{\theta_p, \theta_s} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_p(H_p(f_i), Y_p) + \lambda \mathcal{L}_s(H_s(f_i), Y_s), \tag{3}$$

where N is the total number of images and λ is weight to balance two loss functions. With this method, the parent classification learns a feature space through supervised training via $\mathcal{L}_p$, while the sub-category objective $\mathcal{L}_s$ explores the feature sub-space and provides additional gradients to enhance feature representations f, which is used to compute CAM via (1).

Iterative Optimization. The proposed unsupervised clustering scheme in (2) relies on the feature f to discover sub-category pseudo labels. As such, the learned features via only the objective $\mathcal{L}_p$ could be less discriminative for the clustering purpose. To mitigate this issue, we adopt an iterative training method by alternatively updating (2) and (3). Therefore, features f are first enhanced through the sub-category objective, and in turn facilitate the clustering

Algorithm 1 Learning Sub-category Discovery for CAM

Input: Image I; Parent Label Y_p; Category Number C; Sub-category Number K
Output: Class Activation Map M^c

Model: Feature extractor E; Parent Classifier $(H_p; \theta_p)$; Sub-category Classifier $(H_s; \theta_s)$

Optimize $\{E, H_p\}$ with Y_p via $\mathcal{L}_p$
while Training **do**
  Extract features via $f = E(I)$
  for $c \leftarrow 1$ to C **do**
    Generate pseudo labels Y_s^c with f via (2)
  Optimize $\{E, H_p, H_s\}$ with $\{Y_p, Y_s\}$ via (3)
Compute M^c via (1)

process to generate better pseudo ground truths, which are then used to learn better feature representations in network training. The overall optimization for generating final class activation maps is summarized in Algorithm 1.

3.3. Implementation Details

In this section, we describe implementation details of the proposed framework and the following procedures to produce final semantic segmentation results. All the source code and trained models are available at https://github.com/Juliachang/SC-CAM.

Classification Network. In this work, the ResNet-38 architecture [41] is used for the CAM model, and the training procedure is similar to that in [1]. The network consists of 38 convolution layers with wide channels, followed by a 3×3 convolution layer with 512 channels for better adaptation to the classification task, a global average pooling layer for feature aggregation, and two fully-connected layers for image and sub-category classification, respectively. The model is pre-trained on the ImageNet [10] and is then fine-tuned on the PASCAL VOC 2012 dataset. We use the typical techniques based on the horizontal flip, random cropping, and color jittering operations to augment the training data set. We also randomly scale input images to impose scale invariance in the network.

We implement the proposed framework with PyTorch and train on a single Titan X GPU with 12 GB memory. To train the classification network, we use the Adam optimizer [21] with initial learning rate of 1e-3 and the weight decay of 5e-4. In practice, we use $\lambda = 5$ and $K = 10$ in all the experiments unless specified otherwise. For iterative training, we empirically find that the model converges after training for 3 rounds. In the experimental section, we show studies for the choice of K and iterative training results.

Table 1: Performance comparison in mIoU (%) for evaluating activation maps on the PASCAL VOC training and validation sets.

	Training Set		Validation Set	
Method	CAM	CAM+RW	CAM	CAM+RW
AffinityNet [1]	48.0	58.1	46.8	57.0
Ours	50.9	63.4	49.6	61.2

Semantic Segmentation Generation. Based on the response map generated by our method as in Algorithm 1, we adopt the random walk method via affinity [1] to refine the map as pixel-wise pseudo ground truths for semantic segmentation. In addition, as a common practice, we use dense conditional random fields (CRF) [23] to further refine the response to obtain better object boundaries. To train the segmentation network, we utilize the Deeplab-v2 framework [5] with the ResNet-101 architecture [15] as the backbone model.

4. Experimental Results

In this section, we first present the main results and analysis of the initial response generated by our method. Second, we show the final semantic segmentation performance on the PASCAL VOC dataset [11] against the state-of-the-art approaches. More results can be found in the supplementary material.

4.1. Evaluated Dataset and Metric

We evaluate the proposed approach on the PASCAL VOC 2012 semantic segmentation benchmark [11] which contains 21 categories, including one background class. Each image contains one or multiple object classes. Following previous weakly-supervised semantic segmentation methods, we use augmented 10,528 training images present in [14] along with their image-level labels to train the network. To evaluate the training set, we use the set without augmentation which has 1,464 examples. We adopt 1,449 images in the validation set and 1,456 images in the test set to compare our results with other methods. For all experiments, the mean Intersection-over-Union (mIoU) ratio is used as the evaluation metric. The results for the test set are obtained from the official PASCAL VOC evaluation website.

4.2. Improvement on Initial Response

In Table 1, we show the mean IoU of the segments computed using the CAM on both the training and validation sets. We present results after applying the refinement step to the activation map, i.e., CAM + random walk (CAM + RW). Table 1 shows that our approach significantly improves the IoU over AffinityNet [1] by almost 3% using

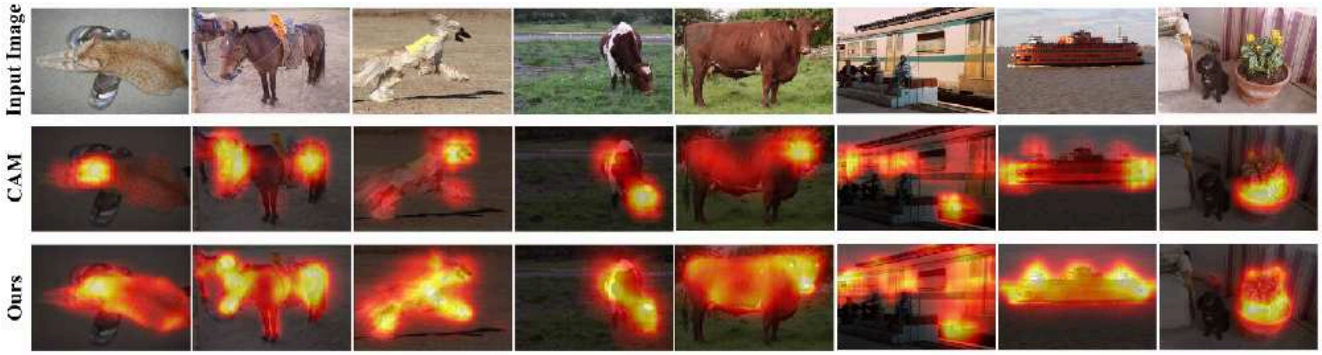

Figure 3: Sample results of initial responses. Our method often generates the response map that covers larger region of the object (i.e., attention on the body of the animal), while the response map produced by CAM [46] tends to highlight small discriminative parts.

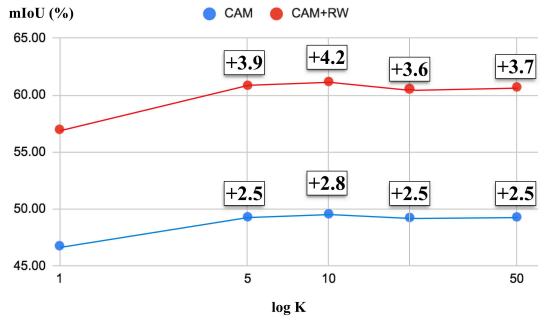

Figure 4: Ablation study for K. We show that the proposed method performs robustly with respect to K and is consistently better than the original CAM that did not apply clustering to discover sub-categories. We mark the value of mIoU of the original CAM at $K = 1$ and the improved mIoUs are presented.

CAM and more than 4% for CAM+RW. The improved initial response maps facilitate the downstream task in generating pixel-wise pseudo ground truths for training the semantic segmentation model.

In Figure 3, we show comparisons of generated CAMs by the conventional classification loss $\mathcal{L}_p$ [46] and the proposed method via sub-category discovery summarized in Algorithm 1. Visual results show that our method is able to localize more complete object regions, while the original CAM only focuses on discriminative object parts. We also note that this is essentially critical for the refinement stage that takes the response map as the input.

4.3. Ablation Study and Analysis

To demonstrate how our method helps improve feature representations and allow the network pay more attention to other object parts via exploiting the sub-category information, we present extensive analysis in this section. Here, all the experimental results are based on the PASCAL VOC validation set.

Table 2: Segmentation quality of the initial response at different rounds of training on the PASCAL VOC 2012 validation set. We show there is a gradual improvement on both mIoU and F-Score metrics.

Round	mIoU (%) ↑	F-Score ↑
#0 (CAM)	46.8	65.1
#1	48.0	65.6
#2	48.7	66.6
#3	**49.6**	**67.0**

Effect of Sub-Category Number K. We first study how the sub-category number K affect the performance of the proposed method. In Figure 4, we use $K = \{5, 10, 20, 50\}$, and show that the proposed method performs robustly with respect to K (within a wide range) and consistently better than the original CAM method (i.e., $K = 1$). The results also validate the necessity and importance of using more sub-categories (i.e., $K > 1$) to generating better response maps. Considering the efficiency and accuracy, we use $K = 10$ for each parent class in all the experiments. As a future work, it is of great interest to develop an adaptive method to determine the sub-category number [33], which can reduce the redundant sub-categories and make the approach more efficient.

Iterative Improvement. To demonstrate the effectiveness of our iterative training process, we show the gradual improvement on the segment quality in Table 2. We present the results of mIoU and F-Score that accounts for both the recall and precision measurements, in which they are important cues to validate whether the activation map is able to cover object parts. Compared to the results in round #0, which is the original CAM, our method gradually improves both metrics as training more rounds.

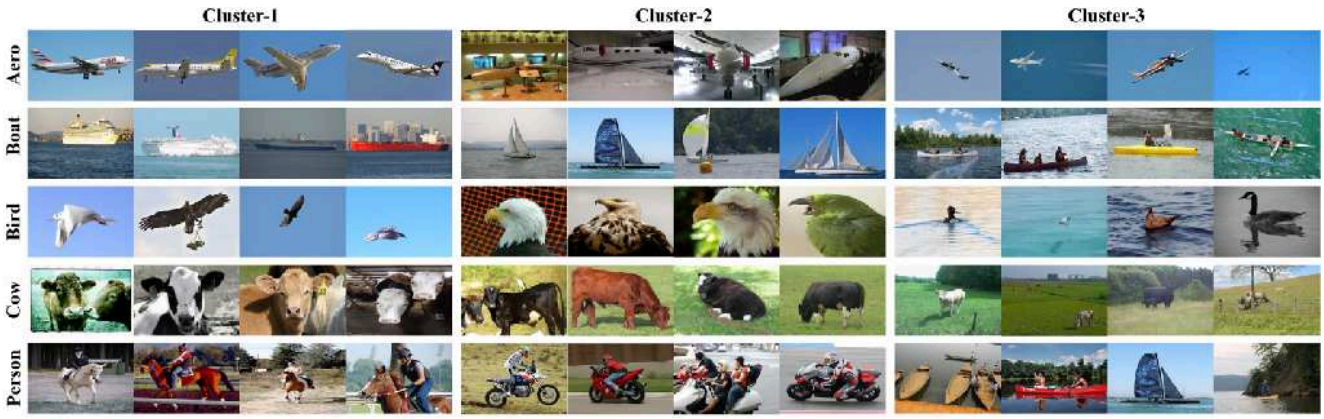

Figure 5: Clustering results of the last round model (#3). We show 3 clusters for each parent class and demonstrate that our learned features are able to cluster objects based on their size (*Aeroplane, Bird, Cow*), context (*Aeroplane, Bird, Person*), type (*Boat, Bird*), pose (*Cow*), and interaction with other categories (*Person*).

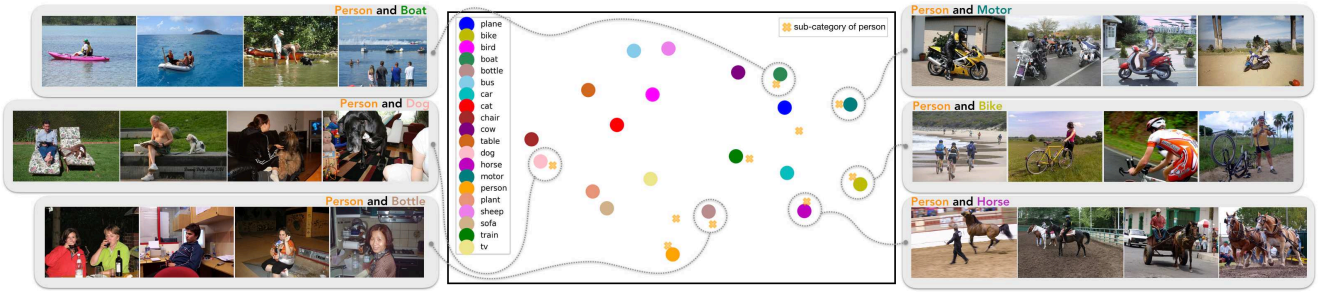

Figure 6: Visualizations of weights based on the t-SNE method that illustrates the relationships on semantic-level between parent classifier and the person sub-category classifier. We show that one person sub-category is usually close to one parent class, as they often co-appear in the same image, as shown in example images on two sides.

Clustering Results. Since the ground truth labels are not available for sub-categories, we present visualizations of clustering results in Figure 5 to measure the quality, in which each parent class shows 3 example clusters. Our method is able to cluster objects based on their size (*Aeroplane, Bird, Cow*), context (*Aeroplane, Bird, Person*), type (*Boat, Bird*), pose (*Cow*), and interaction with other categories (*Person*). For instance, persons with different categories, e.g., horse, motobike, and boat, are clustered into different groups. This visually validates that our learned feature representations are enhanced via the sub-category objective in an unsupervised manner. More visual comparisons are presented in the supplementary material.

Weight Visualization. In order to understand how our learning mechanism improves the clustering quality, we visualize the distribution of the classifier weights, i.e., θ_p and θ_s, via t-SNE [36]. As such, we are able to find the relationship between the parent classifier H_p and the sub-category module H_s. Figure 6 shows the visualization of weights, in which we take the sub-categories of person (denoted as yellow cross symbols) as the example, since the person category has more interactions with other parent classes (denoted as solid circles). It illustrates that one person sub-category is often close to one parent class, e.g., sub-category *person* and parent class *bike*, which makes sense as those two categories usually co-appear in the same image (see example images in Figure 6 on two sides).

4.4. Semantic Segmentation Performance

After generating the pseudo ground truths as the results in Table 1 (i.e., CAM + RW), we use them to train the semantic segmentation network. We first compare our method with recent work using the ResNet-101 backbone or other similarly powerful ones in Table 4. On both validation and testing sets, the proposed algorithm performs favorably against the state-of-the-art approaches. We also note that, most methods focus on improving the refinement stage or network training, while ours improves the initial step to generate better object response maps.

In Table 3, we show detailed results for each category on the validation set. We compare two groups of results

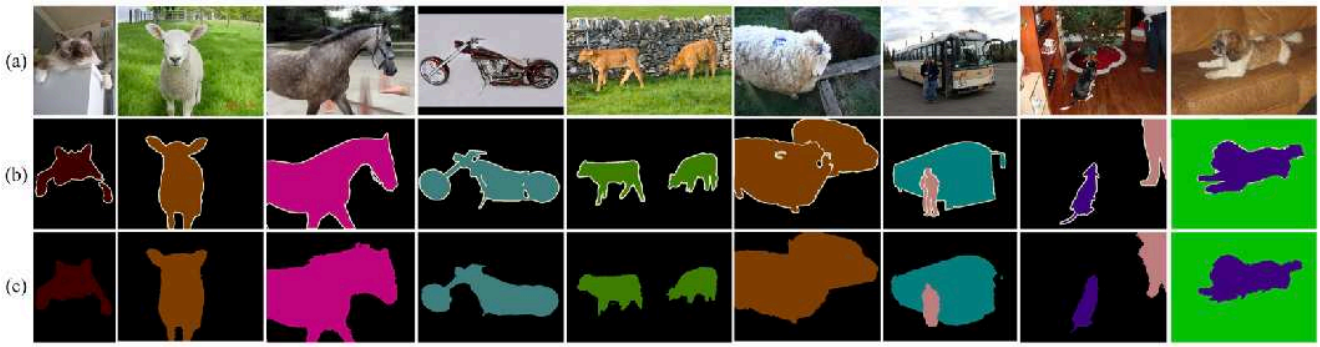

Figure 7: Qualitative results on the PASCAL VOC 2012 validation set. (a) Input images. (b) Ground truth. (c) Our results.

Table 3: Semantic segmentation performance on the PASCAL VOC 2012 validation set. Bottom group contains results with CRF refinement, while the top group is without CRF. Note that 11/20 classes obtain improvements using our approach w/ CRF. The best three results are in red, green and blue, respectively.

Method	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
AffinityNet [1]	88.2	68.2	30.6	81.1	49.6	61.0	77.8	66.1	75.1	29.0	66.0	40.2	80.4	62.0	70.4	73.7	42.5	70.7	42.6	68.1	51.6	61.7
Ours (w/o CRF)	88.1	49.6	30.0	79.8	51.9	74.6	87.7	73.7	85.1	31.0	77.6	53.2	80.3	76.3	69.6	69.7	40.7	75.7	42.6	66.1	58.2	64.8
MCOF [38]	87.0	78.4	29.4	68.0	44.0	67.3	80.3	74.1	82.2	21.1	70.7	28.2	73.2	71.5	67.2	53.0	47.7	74.5	32.4	71.0	45.8	60.3
Zeng et al. [43]	90.0	77.4	37.5	80.7	61.6	67.9	81.8	69.0	83.7	13.6	79.4	23.3	78.0	75.3	71.4	68.1	35.2	78.2	32.5	75.5	48.0	63.3
FickleNet [24]	89.5	76.6	32.6	74.6	51.5	71.1	83.4	74.4	83.6	24.1	73.4	47.4	78.2	74.0	68.8	73.2	47.8	79.9	37.0	57.3	64.6	64.9
Ours (w/ CRF)	88.8	51.6	30.3	82.9	53.0	75.8	88.6	74.8	86.6	32.4	79.9	53.8	82.3	78.5	70.4	71.2	40.2	78.3	42.9	66.8	58.8	66.1

Table 4: Comparison of weakly-supervised semantic segmentation methods on the PASCAL VOC 2012 val and test sets. In addition, we present methods that aim to improve the initial response with ✓ in the "Init. Res." column.

Method	Backbone	Init. Res.	Val	Test
MCOF $_{\text{CVPR'18}}$ [38]	ResNet-101		60.3	61.2
DCSP $_{\text{BMVC'17}}$ [4]	ResNet-101		60.8	61.9
DSRG $_{\text{CVPR'18}}$ [18]	ResNet-101		61.4	63.2
AffinityNet $_{\text{CVPR'18}}$ [1]	Wide ResNet-38		61.7	63.7
SeeNet $_{\text{NIPS'18}}$ [17]	ResNet-101	✓	63.1	62.8
Zeng *et al* $_{\text{ICCV'19}}$ [43]	DenseNet-169		63.3	64.3
BDSSW $_{\text{ECCV'18}}$ [13]	ResNet-101		63.6	64.5
OAA $_{\text{ICCV'19}}$ [19]	ResNet-101	✓	63.9	65.6
CIAN $_{\text{CVPR'19}}$ [12]	ResNet-101		64.1	64.7
FickleNet $_{\text{CVPR'19}}$ [24]	ResNet-101	✓	64.9	65.3
Ours	ResNet101	✓	**66.1**	**65.9**

with (bottom) or without (top) applying the CRF [23] refinement to the final segmentation outputs. Compared to the recent FickleNet [24] method that also focuses on improving the initial response map, the proposed algorithm performs favorably for the segmentation task in terms of the mean IoU. We also note that, our results without applying CRF (mIoU as 64.8%) already achieves similar performance compared with the FickleNet (mIoU as 64.9%). In Figure 7, we present some examples of the final semantic segmentation results, and show that our results are close to the ground truth segmentation.

5. Conclusions

In this paper, we propose a simple yet effective approach to improve the class activation maps by introducing a self-supervised task to discover sub-categories in an unsupervised manner. Without bells and whistles, our approach performs favorably against existing weakly-supervised semantic segmentation methods. Specifically, we develop an iterative learning scheme by running clustering on image features for each parent class and train the classification network on sub-category objectives. Unlike other existing schemes that aggregate multiple response maps, our approach generates better initial predictions without introducing extra complexity or inference time to the model. We conduct extensive experimental analysis to demonstrate the effectiveness of our approach via exploiting the sub-category information. Finally, we show that our algorithm produces better activation maps, thereby improving the final semantic segmentation performance.

Acknowledgments. This work is supported in part by the NSF CAREER Grant #1149783, and gifts from eBay and Google.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, 2018. 1, 2, 3, 4, 5, 8

[2] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *ECCV*, 2016. 1

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018. 3

[4] Arslan Chaudhry, Puneet K Dokania, and Philip HS Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. In *BMVC*, 2017. 8

[5] 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. *CoRR*, abs/1606.00915, 2016. 1, 5

[6] Yi-Wen Chen, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Vostr: Video object segmentation via transferable representations. *International Journal of Computer Vision*, 02 2020. 1

[7] Adam Coates and Andrew Y Ng. Learning feature representations with k-means. In *Neural networks: Tricks of the trade*, pages 561–580. Springer, 2012. 3

[8] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015. 1

[9] Virginia R de Sa. Learning classification with unlabeled data. In *NIPS*, 1994. 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. "http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html". 2, 5

[12] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Cian: Cross-image affinity net for weakly supervised semantic segmentation. In *CVPR*, 2019. 2, 3, 8

[13] Ruochen Fan, Qibin Hou, Ming-Ming Cheng, Gang Yu, Ralph R Martin, and Shi-Min Hu. Associating inter-image salient instances for weakly supervised semantic segmentation. In *ECCV*, 2018. 2, 3, 8

[14] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 5

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[16] Jonathan helhamer, Evan an Long and Trevor Darrell. Fully convolutional networks for semantic segmentation. *TPAMI*, 2016. 1

[17] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *NIPS*, 2018. 2, 8

[18] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *CVPR*, 2018. 1, 2, 3, 8

[19] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In *CVPR*, 2019. 8

[20] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 2017. 1

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[22] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016. 1, 2

[23] Philipp Krahenbuhl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011. 5, 8

[24] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *CVPR*, 2019. 2, 8

[25] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *CVPR*, 2018. 2

[26] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, 2016. 1

[27] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 3

[28] George Papandreou, Liang-Chieh Chen, Kevin Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. In *ICCV*, 2015. 1

[29] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015. 1

[30] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 3

[31] Mattis Paulin, Matthijs Douze, Zaid Harchaoui, Julien Mairal, Florent Perronin, and Cordelia Schmid. Local convolutional features with unsupervised training for image retrieval. In *CVPR*, 2015. 3

[32] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015. 1

[33] Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient parameter-free clustering using first neighbor relations. In *ICCV*, 2019. 6

[34] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 2017. 2

[35] Yi-Hsuan Tsai, Guangyu Zhong, and Ming-Hsuan Yang. Semantic co-segmentation in videos. In *ECCV*, 2016. 1

[36] L. J. P van der Maaten and G. E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 7

[37] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *CVPR*, 2017. 1

[38] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *CVPR*, 2018. 1, 2, 3, 8

[39] 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 *CVPR*, 2017. 2

[40] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *CVPR*, 2018. 1, 2, 3

[41] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019. 5

[42] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 1

[43] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *CVPR*, 2019. 8

[44] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *CVPR*, 2018. 2

[45] Guangyu Zhong, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised video scene co-parsing. In *ACCV*, 2016. 1

[46] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 1, 2, 3, 4, 6