Consistency-CAM: Towards Improved Weakly Supervised Semantic Segmentation

Sai Rajeswar*[1, 2]
rajsai24@gmail.com
Issam Laradji*[2]
issam.laradji@servicenow.com
Pau Rodriguez*[2]
pau.rodriguez@servicenow.com
David Vazquez[2]
david.vazquez@servicenow.com
Aaron Courville[1,3]
aaron.courville@umontreal.ca

[1] Montréal Institute of Learning Algorithms, Université de Montréal.
[2] ServiceNow Research,
[3] CIFAR Fellow.

Abstract

Semantic segmentation is a popular task that has piqued the interest of many industries and research communities. However, acquiring segmentation labels is costly as it often requires carefully annotating the boundaries of the objects of interest. This has triggered research on weakly supervised methods with image-level labels that are less costly to obtain. Existing methods leverage pseudo-labels produced from class activation maps (CAM) generated with models pre-trained on ImageNet. Using CAMs introduces two different challenges. First, ImageNet pre-training biases models to predict a single object per image. Second, pseudo-labels are noisy. In this work, we address the first problem by pre-training the backbone with multi-label iterated learning. In the literature, the second problem is usually alleviated by introducing an additional consistency loss during the backbone pre-training or as an additional CAM refinement step. Here, we propose a generalization of Puzzle-CAMs consistency loss that supports multiple augmentations and tiling resolutions, which helps to further reduce the noise in CAMs and improve the final segmentation performance. The results show improved results in both PASCAL VOC and COCO in the weakly supervised settings compared to existing methods.

1 Introduction

Semantic segmentation is a fundamental task for many computer vision applications, from autonomous driving to medical imaging [10, 43]. To segment an image, a model must assign a class label to each pixel. Thus, good performance depends on a high-level semantic understanding of the image's composition and contents, as well as fine-grained attention to low-level pixel details. This is typically achieved by fine-tuning a large pre-trained image model on pixel-level annotations [10, 18].

*Equal Contribution. Correspondence to <rajsai24@gmail.com>

Obtaining pixel-level annotations, however, can be significantly more expensive than image-level labels. Namely, assigning a label to every pixel can be labor-intensive and time-consuming. This has raised interest in using weaker forms of supervision such as image-level annotations, scribbles, point annotations, or bounding boxes [41, 48, 51, 53]. The image-level class label is the most accessible form of annotation since it already exists in large-scale datasets like ImageNet [15]. Thus in this work, we focus on weakly supervised semantic segmentation (WSSS) using only image-level class labels.

A particularly successful WSSS approach is to train a model for supervised image classification and produce pixel-level pseudo-labels with class activation maps (CAM) [56] and then train a segmentation network on those pseudo-labels [1, 24]. Critically, the success of WSSS relies on the accuracy of the pseudo annotations. However, one of the problems with using pseudo-labels is that they tend to be noisy. This noise is usually alleviated with post-processing steps like conditional random fields [9, 36] or inductive biases during training, such as a consistency loss [9]. Puzzle-CAM [31] is a recent example that obtained state-of-the-art performance by leveraging both noise reduction methods. One of the main contributions of Puzzle-CAM is a novel consistency loss that tiles the input in a grid to produce a CAM for each tile. Then it optimizes the consistency between the CAM resulting from stitching back all the tiles and the CAM produced by the original untiled input.

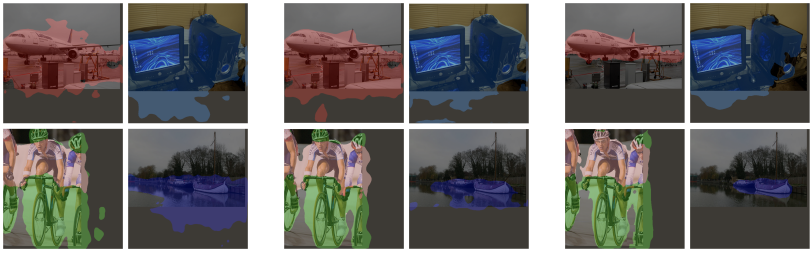
In this work, we identify three different improvements to increase CAM-based WSSS methods' performance without increasing inference time. First, it is known that ImageNet pre-trained backbones are biased toward predicting a single class per image [6]. Since semantic segmentation datasets often contain many distinct objects, this bias could hinder final performance. Thus, we pre-train the backbone with a modified version of multi-label iterated learning (MILe) [46], an iterated learning procedure that allows ImageNet models to produce multiple labels per image from single label annotations. Second, state-of-the-art methods like Puzzle-CAM output a fixed number of tiles per input image. This constrains the consistency loss to optimize the model for that particular number of tiles. We relax this by randomizing the number of tiles at every iteration. Finally, we generalize the tiling operation to a larger set of transformations. The resulting method, Consistency-CAM achieves 68.2% mIoU on PASCAL VOC and 40.8% mIoU on COCO datasets.

Our contributions are: (i) We identify three different improvements for CAM-based WSSS methods: multi-label relaxation, variable tile size, and consistency with random augmentations. (ii) We show that each improvement individually increases mIoU on different backbones. (iii) The resulting model, which we name Consistency-CAM, outperforms existing methods with the same level of supervision on the PASCAL VOC 2012 dataset and demonstrate competitive performance on the COCO dataset.

2 Related Work

Our work lies at the intersection of weakly supervised, self-supervision and iterated learning. Below we describe each of these topics and how existing methods relate to our algorithm.

Weakly Supervised Semantic Segmentation (WSSS) is a popular topic where the idea is to vastly reduce the required annotation cost for acquiring a training set. According to Bearman et al. [5], manually collecting image-level and point-level labels for the PASCAL VOC dataset [18] took only 20.0 and 22.1 seconds per image, respectively. These annotation methods are an order of magnitude faster than acquiring full segmentation labels, which is



(a) **Puzzle-CAM** (b) **Puzzle-CAM+aug** (c) **Consistency-CAM**

Figure 1: Qualitative visualizations of predicted segmentation masks on VOC 2012.

239.0 seconds. Other forms of weaker labels were explored, such as bounding boxes [32] and image-level annotation [57]. Li et al. [40] have shown that using a denoised version of pseudo-masks generated by class activation maps has shown to be effective for both COCO and PASCAL datasets. An affinity-based denoising method of those pseudo masks was proposed by Ahn and Kwak [2]. The goal is to use a network that can learn to refine masks based on neighboring features and colors. Another method to denoise those masks is recursive learning and data augmentation combining different masks [30]. Our work is inspired by Puzzle-CAM [31], which uses a reconstruction regularization by breaking the image into smaller patches and enforce a consistency between the segmentation predictions of the image and corresponding smaller patches. In this work, we focus on image-level supervision.

Self-supervised Learning (SSL) approaches aim at designing effective surrogate tasks to train a model without additional extrinsic annotations. Some self-supervised pretext strategies are relative affine and spatial transformation prediction [16, 20], inpainting [44], colorization [37] and recently, contrastive learning over multiple augmentations that assume visual embeddings are invariant under a set of transformations [9, 11]. In segmentation tasks, SSL based methods attempt to recognize classes of dense pixels. Clustering is one way to discover semantic classes and perform recognition on pixels or patches. IIC [29] performs invariant information clustering by maximizing the mutual information between encoded image pairs. More recently, Cho et al. [12] conducts alternative offline pixel-wise clustering and online training, where the training is led by the invariance and equivariance objective on the assigned clusters. Puzzle-CAM [31] breaks the image into tiles and ensures that the global segmentation output on the whole image matches the individual tiles. Our approach generalizes Puzzle-CAM using augmentation invariant SSL framework with more general transformations and by incorporating iterated learning.

Iterated Learning was proposed by Kirby [33, 34] to model language evolution via cultural transmission in humans. Languages need to be expressive and compressible to be effectively transmitted through generations. This learning bottleneck favors languages that are compositional as they can be easily learned by the offsprings and support generalization. It has seen many successful applications, especially in the emergent communication literature [13, 14, 21, 47]. Recently, MILe [46] emerged as an iterated learning procedure that allows ImageNet models to produce multiple labels per image from single label annotations. Our model uses a variation of MILe to help improve its WSSS capabilities.

3 Methodology

We follow the typical multi-stage WSSS pipeline based on CAM [52]. First, a model is pre-trained for image classification. Second, the pre-trained model extracts class activation maps (CAM) with a refinement process to reduce noise [31]. Finally, CAMs are used as pixel-level pseudo-labels to train the segmentation model. In this work, we focus on the first two stages.

3.1 Model pretraining.

ImageNet pre-trained models tend to predict a single label per image. This constraint is unrealistic for semantic segmentation, where multiple objects can be present at the same time. Following Rajeswar et al. [46], we leverage multi-label iterated learning to learn a multi-label backbone from single labels.

Multi-label Prediction via Iterated Learning. Following the same procedure described in [46], we train a teacher network with binary cross entropy for a few (k_t) iterations. Then we initialize a student network with the teacher weights and train it for a few iterations (k_s) on pseudo-labels predicted by the teacher. Pseudo-labels are obtained by applying a threshold (denoted by ρ) on the teacher's output, resulting in a binary vector. Once trained, the student becomes the teacher, and the cycle is repeated until convergence. We restrict the imitation learning phase to a limited learning budget (an essential component of the iterated learning framework [33]). This learning bottleneck regularizes the student model to avoid the specific irregularities in the data. In our setting, we enforce the bottleneck via the number of learning updates akin to Rajeswar et al. [46].

Improving GAP with noisy-or. Typically, CAMs are obtained from a model trained by applying global average pooling (GAP) to the backbone's feature maps before applying a classifier that returns a 1-d vector of class probabilities [56]. However, this forces the majority of pixels in a feature map to belong to a certain class so that the probability for that class is high when they are averaged and fed to a classifier. In images where there are multiple objects, this would be equivalent to enforcing the presence of an object in the whole image. To alleviate this problem, we propose applying the classifier before GAP, obtaining class activation maps (CAMs), and then reducing the CAMs to a 1-d vector of class probabilities by applying the noisy-or operation. The noisy-or operation is a function originally designed for Bayesian networks [45] that assumes a disjunctive interaction among the parents of a node (e.g., multiple class predictions for a single image). The advantage of the noisy-or is that just one pixel must belong to a class to mark it as present in the output. Formally, let d be the spatial dimensionality of a CAM (height times width), and c the number of classes in the dataset. The noisy-or function $f : \mathbb{R}^{d,c} \rightarrow \mathbb{R}^c$ returns a vector of class probabilities when applied to the CAM (A): $p(class_i = 1) = 1 - \prod_{j=1}^{d}(1 - A_{i,j})$. We perform this operation both for the teacher and the student training during iterated learning (see Fig.4).

3.2 CAM Refinement with Puzzle-CAM.

CAMs obtained after training a model for image classification tend to be noisy. The recent Puzzle-CAM proposes to fine-tune the pre-trained model with a consistency loss that ensures that CAMs produced in local regions of an image match the overall CAM of the image (A). For that, they divide the image (I) in 2×2 equally-sized tiles ($I_{1,1}...I_{2,2}$). Then, the CAMs

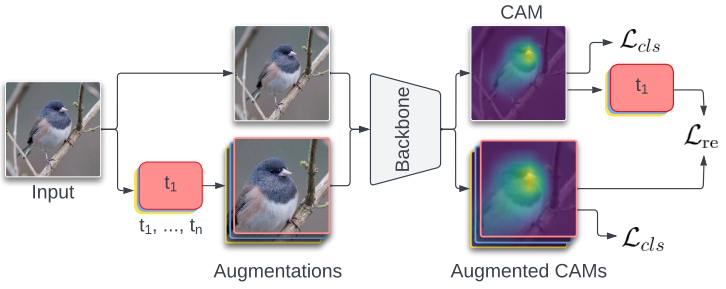


Figure 2: **Overview of Consistency-CAM.** We train a model to produce CAMs from an image and multiple augmentations of it ($t_1..t_n$). Then we perform the same augmentations to the CAM of the original image and enforce them to be the same as the CAMs of the augmentations. An additional classification loss is applied to each of the CAMs.

of those tiles ($A_{1,1}...A_{2,2}$) are merged back into a single CAM (A^{re}). The consistency loss encourages the reconstructed CAM to match the original CAM. Their overall objective is:

$$\mathcal{L}_{\text{puzzle}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{p-cls}} + \alpha\mathcal{L}_{\text{re}}, \tag{1}$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{p-cls}$ is the multi-label soft-margin loss with respect to the global average pooling (GAP) of A and A^{re}, respectively. The scalar α weighs the consistency loss (details outlined in Sec. 4.1). $\mathcal{L}_{re}$ is the consistency loss:

$$\mathcal{L}_{\text{re}} = ||A - A^{re}||_1. \tag{2}$$

The goal is to encourage tile-based predictions to be consistent with the overall image prediction. To satisfy this, the model must reduce noise in the generated CAMs.

We propose to generalize the Puzzle-CAM objective in two different ways. First, we use multi-resolution tiling consistency, where we relax the tiling operation from 2×2 to any value in $\{2 \times 2, 4 \times 4, 8 \times 8, 16 \times 16\}$. Second, we introduce additional consistency losses that enforce CAM to be invariant to different data augmentations (see Fig.2). Since CAMs generated from augmented images cannot directly be compared with the original CAM, we apply the same augmentations to the original CAM before computing the consistency loss. For example, given the CAM of a flipped and an unflipped version of an image, we would flip the CAM of the unflipped version before comparing it with the CAM of the flipped version. Formally, let A^t be the CAM corresponding to an image (I) that has been transformed with transformation $A^t = t(A)$. The consistency loss for augmented images is defined as $||A^t - t(A)||_1$. The overall loss for augmentation is the sum of its consistency loss and classification loss (e.g. Eq. 1). We apply this loss on three different augmentations: (i) random crop ($\mathcal{L}_{rc}$), (ii) horizontal flip ($\mathcal{L}_{hf}$), (iii) random resize ($\mathcal{L}_{rs}$). The final loss that we optimize during training is:

$$\mathcal{L} = \mathcal{L}_{\text{puzzle}} + \mathcal{L}_{rc} + \mathcal{L}_{hf} + \mathcal{L}_{rs}. \tag{3}$$

4 Experiments

4.1 Experimental Setup

Implementation Details. For learning CAMs and pseudo-labels, we used the stochastic gradient descent (SGD) optimizer with weight decay $4e-5$. The initial learning rate was

Model	Backbone	Flops	mIoU	
			Base	NoisyOR
DeepLabV3	ResNet50	51.4G	76.9	78.0
DeepLabV3	ResNet101	72.1G	77.3	78.5
DeepLabV3+	ResNet50	62.7G	77.2	78.3
DeepLabV3+	ResNet101	83.4G	78.3	79.0

Figure 3: **Performance on Pascal VOC2012** using our Pre-trained model.

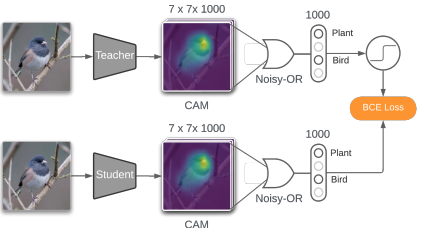


Figure 4: **Backbone pre-training** with Noisy-OR.

set to 0.1 and decayed polynomially at a rate of 0.9. The images were randomly re-scaled in the range of [320, 640] and then cropped to 512 × 512 as the inputs to the model. For the weight on reconstruction term in Eq.1, we used $\alpha = 4$ as the maximum value. Furthermore, the α value is linearly increased up to its maximum value by half epochs for all the experiments. The pipeline and the details are similar to the ones followed in [3, 31]. After obtaining the final pseudo-labels for segmentation, we train the DeepLabv3+ model with ResNet backbones for the segmentation. We followed the same procedure and details described in [8]. A polynomial schedule with an initial learning rate of 0.007 was used, batch normalization parameters [28] were fine-tuned when output stride = 16, and used random scale data augmentation during training. When training the COCO 2014 dataset, we used the hyperparameters identical to the ones used for PASCAL VOC 2012 dataset.

Datasets. We evaluate our approach on the PASCAL VOC2012 and the COCO datasets. PASCAL VOC consists of 20 foreground object classes and one background class. We augment this version of the dataset as we leverage extra annotations provided by the Semantic Boundary Dataset [22], resulting in 10,582 training images. Note that the original dataset contains 1,464 (train), 1,449 (val), and 1,456 (test) pixel-level annotated images. During the whole training process, we only adopt the image-level class labels for supervision. COCO consists of 80 categories belonging to a wide variety of everyday objects. The train set is composed of 80K images, while the validation set has 40137 images. We follow the experimental setup of [24, 25].

Metrics. Following common practice for semantic segmentation [2, 18], we evaluate our models with the Intersection over Union (IoU) which measures the overlap between the prediction and the ground truth: IoU = TP / TP +FP +FN , where TP, FP, and FN are the number of true positive, false positive, and false negative pixels across all images in the test set. Since each image may contain multi-class labels, we calculate the mean intersection over union (mIoU) of all classes to evaluate the performance.

4.1.1 Methods and Baselines

We compare our method with **Puzzle-CAM** [31] and the following methods:
AffinityNet [1] refines CAMs by leveraging an affinity network that predicts the semantic affinity between neighboring pairs of pixels.
SEAM [34] refines CAMs using a pixel correlation module that captures context appearance information for each pixel and alters the original CAMs by using learned affinity maps.
IRNet [3] leverages class-boundary maps to learn pairwise affinity scores to refine instance-wise CAMs for weakly supervised instance segmentation.

Method	Backbone	Supervision	val mIoU	test mIoU
AffinityNet [1]	ResNet50	$\mathcal{I}$	61.7	63.7
DSRG [27]	ResNet101	$\mathcal{I}+\mathcal{S}$	61.3	63.2
SeeNet [26]	ResNet101	$\mathcal{I}+\mathcal{S}$	63.1	64.3
IRNet [3]	ResNet50	$\mathcal{I}$	63.5	64.8
Puzzle-CAM [31]	ResNet50	$\mathcal{I}$	63.33	63.9
ICD [19]	ResNet101	$\mathcal{I}$	64.1	64.3
SEAM [54]	ResNet38	$\mathcal{I}$	64.5	65.7
FickleNet [38]	ResNet50	$\mathcal{I}+\mathcal{S}$	64.9	65.3
CONTA [17]	ResNet38	$\mathcal{I}$	66.1	66.7
SC-CAM [2]	ResNet101	$\mathcal{I}$	66.1	65.9
Sun et al. [52]	ResNet101	$\mathcal{I}$	66.2	66.9
PMM [39]	ScaleNet101	$\mathcal{I}$	67.1	67.7
Puzzle-CAM	ResNest101	$\mathcal{I}$	66.81	67.7
Consistency-CAM	ResNet50	$\mathcal{I}$	**64.26**	64.4
Consistency-CAM w/o IL	ResNest101	$\mathcal{I}$	68.20	68.5
Consistency-CAM	ResNest101	$\mathcal{I}$	**68.89**	**69.1**

Table 1: **Performance on PascalVOC.** Comparison with existing methods on PASCAL VOC2012 val set. All results are evaluated in mIoU(%). $\mathcal{I}$ represents the image-level label and $\mathcal{S}$ indicates the saliency label.

SEC [35] optimizes a loss that seeds segmentation masks with weak localization cues, another that expands objects based on the information about which classes can occur in an image, and another that constrains the segmentations to coincide with object boundaries.
PMM [39] identifies the category independently and leverages variation smoothing to refine the CAM by their distribution statistics.
CONTA [17] leverages iterative post-processing to refine CAM where it iterates through the whole process of WSSS including a sequence of model training and inference.

4.2 Results

We provide experiments showing the effects of our proposed approach on semantic segmentation tasks. Before diving into the WSSS setup, we study the benefits of our iterated learning procedure for semantic segmentation on PascalVOC. We employ ImageNet [15] pre-trained ResNet-101 [23] that is trained using our aforementioned iterated learning procedure to extract dense feature maps by atrous convolution. The resulting model is then evaluated on the PASCAL VOC 2012 semantic segmentation benchmark in Table 3. We observe that our procedure surpasses baseline methods on both ResNet50 and ResNet101 backbones and with both variants of DeepLab architectures. With DeepLabV3, we observe a substantial improvement on mIOU for ResNet-50 and ResNet-101 respectively than on DeepLabV3Plus. Encouraged by the benefits of Consistency-CAM on the supervised segmentation task, we continue to leverage the pre-training strategy for the WSSS tasks. In Sec. 4.3, we explore the benefits of our complete approach to WSSS tasks on VOC and COCO datasets.

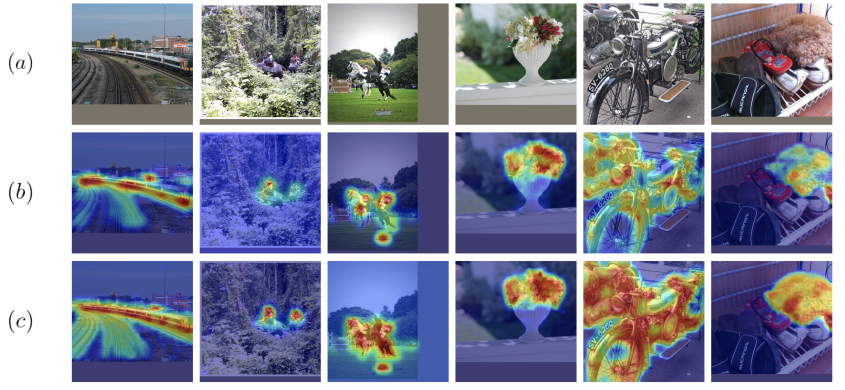


Figure 5: **CAM visualizations on COCO (train).** (a) original images. (b) Puzzle-CAM [31]. (c) Consistency-CAM (ours)

4.3 Weakly Supervised Segmentation

Pseudo-labels from the CAMs are obtained using the combination of losses formulated in Eq. 2. To improve the performance of pixel-level pseudo-labels we also train AffinityNet [2] similar to Jo and Yu [31]. A detailed analysis of the initial accuracy of pseudo-labels on the VOC training set is discussed in a later section (see Sec. 4.4). We train the segmentation model DeepLabV3 [10] using Consistency-CAM on the obtained pseudo-labels of the training set and report the results on PASCAL VOC2012 validation. Table 1 shows that for both ResNet50 and ResNet101, our results outperform those of the existing methods with supervision at the same level (i.e., WSSS-Image Level supervision) or higher granular supervision cues (e.g., salient object supervision requiring auxiliary object boundary information, extra dataset and segment-based object proposals). Specifically, our ResNest101 model surpasses Puzzle-CAM's performance by more than 2% leveraging both the transformations and Iterated Learning (IL) and 1.4% with transformation alone, thereby demonstrating the individual necessity and effectiveness of IL and the consistency loss. The backbones leveraged for the segmentation framework are listed in Table 1. Qualitative results in Fig.1 show that adding additional augmentations to Puzzle-CAM (a) results in more accurate masks (b), and that these masks can be further improved with iterated learning and multi-resolution tiling (c).

4.3.1 WSSS Experiments on COCO

COCO is a more challenging setting than VOC due to the drastic variability in objects size. In particular, there are a significant number of images with tiny objects that makes the task especially difficult with image-level annotations. For this reason our performance is competitive to the ScaleNet101 backbone version of PMM [39] which is more robust to scale changes than ResNet or ResNest backbones. However, our method outperforms Puzzle-CAM as seen quantitatively in Table 2. We also compare with methods that rely on additional granular cues such as the saliency model and segment-based object proposals that have an advantage over our method that use image level supervision. Taken as a whole, these results suggest that our approach is able to effectively leverage weak supervision in the form of high-level annotations. Qualitative visualization of the prediction CAMs and comparisons with the baseline is shown in Fig.5

Method	Backbone	RW	val mIoU
BFBP [49]	VGG16	$\mathcal{I}$	20.4
SEC [35]	VGG16	$\mathcal{I}$	22.4
DSRG [27]	ResNet101	$\mathcal{I}+\mathcal{S}$	26.0
SEAM [52]	ResNet38	$\mathcal{I}$	31.7
CONTA [47]	ResNet38	$\mathcal{I}$	32.8
PMM [39]	Res2Net101	$\mathcal{I}$	35.7
PMM [39]	ScaleNet	$\mathcal{I}$	40.2
Puzzle-CAM	ResNest101	$\mathcal{I}$	38.9
Consistency-CAM	ResNest101	$\mathcal{I}$	**40.8**

Table 2: **Performance on COCO.** Comparison with existing methods on COCO2014 val set. All results are evaluated in mIoU(%). $\mathcal{I}$ represents the image-level label and $\mathcal{S}$ indicates the saliency label.

Method	L_{puzzle}	L_{trans}	$L_{multi-res}$	TrainmIoU
AffiniyNet [1]	-	-	-	47.82
IRNet [3]	-	-	-	48.3
CONTA [47]	-	-	-	48.8
Puzzle-CAM [31]	✓			50.14
Consistency-CAM	✓		✓	51.21
Consistency-CAM	✓	✓		51.88
Consistency-CAM	✓	✓	✓	53.64

Table 3: **Ablation on CAM classification** Comparison with different effects of each component of our method.

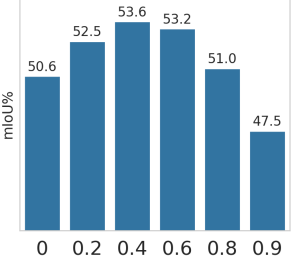


Figure 6: Effect of the IL threshold on performance.

4.4 Ablation Analysis on CAMS and Pseudo-labels

The proposed scheme aims to provide segmentation-specific CAMs to improve the quality of the pseudo-labels. In order to verify the effectiveness of our method in generating CAMs and Pseudo-labels, we summarize the results of the CAMs and the pseudo-labels of the PASCAL VOC2012 training set with competitive methods and analyze different components of our approach. Table 3 shows the initial accuracy of the pseudo-labels on the VOC training set before training the AffinityNet on the obtained pseudo-labels. We observe that our framework achieves the mIoU of 53.6%. Our method surpasses the advanced method Puzzle-CAM by 3% and outperforms CONTA [47]. The Consistency-CAM ablations reveal that both the transformation based consistency loss and the multi-resolution tiling contribute to the improved performance. In a furhter ablation, we study the effect of the threshold used in the iterated learning pre-training on the final performance on PASCAL dataset (see Fig.6). Threshold value (ρ) is used by IL to produce multi-pseudo-labels from sigmoid output activations. Larger threshold values yield lower performance as the student network is constrained to predict sparser labels during the pre-training stage and can potentially force the teacher network to output empty labels.

5 Conclusion

We introduced Consistency-CAM, a WSSS method that enhances Puzzle-CAM during backbone pre-training and CAM refinement. During backbone pre-training, we replace the single-label classification task for multi-label classification with multi-label iterated learning. We also replaced global average pooling by a noisy-or operation which removes the bias that all output pixels in a feature map must belong to the target class. We found that the resulting backbone is more suitable for image segmentation, where multiple objects are typically present in a single image. During CAM refinement we relaxed the Puzzle-CAM consistency loss with multi-resolution tiles and three additional augmentations. We found that the resulting model outperforms or matches previous performance results on PASCAL VOC and COCO.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4981–4990, 2018.

[3] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[4] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. *arXiv preprint arXiv:2204.07141*, 2022.

[5] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. *European Conference on Computer Vision (ICCV)*, 2016.

[6] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

[7] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[9] 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, 2018. doi: 10.1109/TPAMI.2017.2699184.

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020.

[12] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16794–16804, June 2021.

[13] Michael Cogswell, Jiasen Lu, Stefan Lee, Devi Parikh, and Dhruv Batra. Emergence of compositional language with deep generational transmission. *arXiv preprint arXiv:1904.09067*, 2019.

[14] Gautier Dagan, Dieuwke Hupkes, and Elia Bruni. Co-evolution of language and agents in referential games. *arXiv preprint arXiv:2001.03361*, 2020.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[16] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[17] Zhang Dong, Zhang Hanwang, Tang Jinhui, Hua Xiansheng, and Sun Qianru. Causal intervention for weakly supervised semantic segmentation. In *NeurIPS*, 2020.

[18] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal on Computer Vision (IJCV)*, 2010.

[19] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[20] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

[21] Shangmin Guo, Yi Ren, Serhii Havrylov, Stella Frank, Ivan Titov, and Kenny Smith. The emergence of compositional languages for numeric concepts through iterated learning in neural agents, 2019.

[22] Bharath Hariharan, Pablo Arbelaez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *International Conference on Computer Vision (ICCV)*, 2011.

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

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *ICCV*, 2017.

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[26] Qibin Hou, Peng-Tao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *NeurIPS*, 2018.

[27] Zilong Huang, JiasiWang XinggangWang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[28] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[29] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9865–9874, 2019.

[30] Sang Hyun Jo, In Jae Yu, and Kyung-Su Kim. Recurseed and certainmix for weakly supervised semantic segmentation. *arXiv preprint arXiv:2204.06754*, 2022.

[31] Sanghyun Jo and In-Jae Yu. Puzzle-cam: Improved localization via matching partial and full features. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 639–643. IEEE, 2021.

[32] Anna Khoreva, Rodrigo Benenson, Jan Hendrik Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[33] Simon Kirby. Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation, 5(2):102–110*, 2001.

[34] Simon Kirby. Natural language from artificial life. *Artificial life, 8(2): 185–215*, 2002.

[35] Alexander Kolesnikov and Christoph H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision (ECCV)*. Springer, 2016.

[36] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[37] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European Conference on Computer Vision (ECCV)*, 2016.

[38] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[39] Yi Li, Zhanghui Kuang, Liyang Liu, Yimin Chen, and Wayne Zhang. Pseudo-mask matters in weakly-supervised semantic segmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6944–6953, 2021.

[40] Yi Li, Zhanghui Kuang, Liyang Liu, Yimin Chen, and Wayne Zhang. Pseudo-mask matters in weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6964–6973, 2021.

[41] Yun Liu, Yu-Huan Wu, Peisong Wen, Yujun Shi, Yu Qiu, and Ming-Ming Cheng. Leveraging instance-, image- and dataset-level information for weakly supervised instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. doi: 10.1109/TPAMI.2020.3023152.

[42] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[43] A. A. Novikov, Dimitrios Lenis, David Major, Jiří Hladůvka, Maria Wimmer, and Katja Bühler. Fully convolutional architectures for multiclass segmentation in chest radiographs. *IEEE Transactions on Medical Imaging*, 37:1865–1876, 2018.

[44] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[45] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

[46] Sai Rajeswar, Pau Rodriguez, Soumye Singhal, David Vazquez, and Aaron Courville. Multi-label iterated learning for image classification with label ambiguity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4783–4793, 2022.

[47] Yi Ren, Shangmin Guo, Matthieu Labeau, Shay B. Cohen, and Simon Kirby. Compositional languages emerge in a neural iterated learning model. In *ICLR*, 2020.

[48] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Alexander G. Schwing, and Jan Kautz. Ufo^2 : *Aunifiedframeworktowardsomni* − *supervisedobjectdetection*. *InECCV*, 2020.

[49] Fatemeh Sadat Saleh, Mohammad Sadegh Ali Akbarian, Mathieu Salzmann, Lars Petersson, Stephen Gould, and José Manuel Álvarez. Built-in foreground/background prior for weakly-supervised semantic segmentation. *ArXiv*, abs/1609.00446, 2016.

50] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. 2015.

51] Shih-Po Lee Shun-Yi Pan, Cheng-You Lu and Wen-Hsiao Pen. Weakly-supervised image semantic segmentation using graph convolutional networks. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2021.

52] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *ECCV*, 2020.

53] Pei Wang, Zhaowei Cai, Hao Yang, Gurumurthy Swaminathan, Nuno Vasconcelos, Bernt Schiele, and Stefano Soatto. Omni-detr: Omni-supervised object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9367–9376, June 2022.

54] 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.

55] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *CVPR*, 2014.

56] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

57] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

	bckg	*aero*	*bike*	*bird*	*boat*	*bottle*	*bus*	*car*	*cat*	*chair*	*cow*	*table*	*dog*	*horse*	*mbk*	*person*	*plant*	*sheep*	*sofa*	*train*	*tv*	mIoU
AffinityNet []	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
SEAM []	88.8	68.5	33.3	**85.7**	40.4	67.3	78.9	76.3	81.9	29.1	75.5	48.1	79.9	73.8	71.4	75.2	**48.9**	79.8	40.9	58.2	53.0	64.5
FickleNet []	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
CONTA []	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
ConsisstencyCAM	**90.1**	**81.9**	**35.4**	84.7	**67.6**	67.9	87.5	**80.5**	86.5	31.4	73.9	52.5	**84.0**	74.9	74.6	79.0	44.7	84.1	47.0	78.4	46.6	68.9

Table 4: per-class Categorical semantic segmentation performance on PASCAL VOC 2012 emphval.

6 Appendix