Ensemble of CAM, ECS-CAM & CCAM for Stronger Weakly-Supervised Segmentation

Jiakang Chen

UCL

Abstract. Weakly supervised semantic segmentation (WSSS) reduces the cost of pixel-level annotation by training models with only image-level labels. In this work, we first establish a unified benchmark evaluating three CAM-based localization methods—CAM, Erased CAM (ECS-CAM), and Contrastive CAM (CCAM)—under identical training and augmentation protocols. Building on their complementary strengths, we propose an ensemble framework that fuses multi-method activation maps and applies a lightweight CRF-based refinement to produce high-quality pseudo-pixel labels. To further improve mask completeness and background suppression, we introduce two novel training strategies: (1) augmenting with background-only images by adding a "bg" class, and (2) a pet-specific contrastive fine-tuning stage that treats each breed as a separate category within CCAM. Finally, we use these pseudo-labels to train a U-Net segmentation network and compare its performance against a fully supervised U-Net baseline. Our best weakly supervised model achieves 61.6 %(finetune) and 76 %(pretrain) IoU comparing to the 68.25 % on fully supervised baseline model.

1 Introduction

Deep convolutional neural networks have driven remarkable advances in computer vision tasks such as object detection and semantic segmentation. However, these successes depend critically on large-scale datasets with precise, pixel-level annotations—annotations that are both labor-intensive and costly to acquire at scale. Weakly supervised semantic segmentation (WSSS) seeks to alleviate this bottleneck by training segmentation models using only weak labels (e.g., image-level tags, bounding boxes, scribbles), thereby reducing annotation cost while still achieving competitive performance.

A popular WSSS paradigm leverages class activation maps (CAMs) [6] to localize the most discriminative object regions from image-level labels. While CAMs provide a useful starting point, they typically highlight only the core object parts and fail to cover full object extents, limiting segmentation quality. To address this, Erased Class Activation Mapping (ECS-CAM) [4] iteratively removes highly activated regions, compelling the network to discover additional object parts. More recently, Class-Agnostic Activation Mapping (CCAM) [5] employs contrastive learning to better separate foreground and background, yielding more complete activation masks.

In this work, we build upon these CAM-based techniques and make the following contributions. We first benchmark CAM, ECS-CAM and CCAM under a unified training protocol, then introduce an ensemble of their activation maps plus lightweight post-processing, enhanced by background-only augmentation and class-specific contrastive refinement, to produce high-quality pseudo-labels from image-level tags. Finally, we apply these pseudo-labels to train a segmentation network and compare its performance against a fully supervised counterpart.

The remainder of this paper is organized as follows. In Section 2, we introduce methods on CAM enhancements and pipeline. Section 3 describes our experiments setup and training strategies. Experimental results and ablations are presented in Section 4 and discussed in Section 5, and we conclude in Section 6.

2 Methods

As illustrated in Fig. 1, our framework consists of three stages: (1) **Feature Extraction**, where input images are passed through a modified ResNet-50 backbone [5] and the last two convolutional blocks are retained as deep feature maps; (2) **Pseudo-Label Generation**, in which class activation mapping (CAM), its contrastive learning refinement (CCAM), and Erasing (ECS) produce candidate masks from the feature maps, and the best mask is further refined by a CRF filter[1]; and (3) **Segmentation Training**, where the CRF-filtered mask serves as a pseudo-ground truth to train a U-Net for the final pixel-level segmentation.

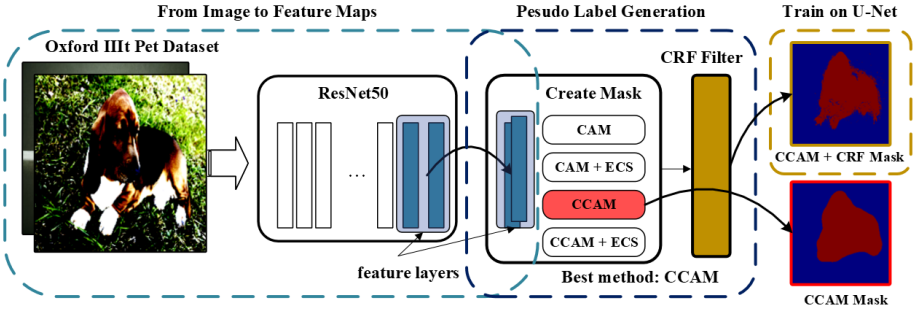


Fig. 1: Overall architecture of the proposed weakly-supervised segmentation pipeline. Deep features are extracted by a modified ResNet-50, converted into class-specific activation maps (CAM, CCAM, ECS), refined via CRF, and finally used as pseudo-labels to train a U-Net.

[1] Due to length limitations, the detailed introduction on CRF [2] is shown at Appendix A

2.1 Class Activation Maps

Class Activation Mapping (CAM) [6] remains a cornerstone of weakly-supervised segmentation, since it produces class-specific heatmaps without requiring pixel-level labels. Given a CNN's final convolutional feature maps $f_k(x, y)$, we form channel-wise descriptors by global average pooling,

$$F^k = \sum_{x,y} f_k(x, y)\,, \tag{1}$$

and obtain the pre-activation score for class c via

$$S_c = \sum_k w_k^c\, F^k = \sum_k w_k^c \sum_{x,y} f_k(x, y)\,, \tag{2}$$

where w_k^c are the weights of the classification layer. Reordering yields the activation map

$$M_c(x, y) = \sum_k w_k^c\, f_k(x, y)\,, \tag{3}$$

which highlights regions driving the class decision. However, CAM typically focuses on only the most discriminative object regions rather than the full object extent. To address this limitation, we apply the following refinement steps.

2.2 Erased Class Activation Mapping

Erased Class Activation Mapping (ECS-CAM) [4] builds on CAM by forcing the model to discover less salient object parts. Specifically, after computing the initial heatmap $M_c(x, y)$ for class c, we suppress its top-activated regions in the input (e.g. via blurring). Re-applying CAM to this erased image yields a secondary map $M_{\text{ecs}}(x, y)$ that highlights previously neglected areas. We then fuse both maps:

$$M_{\text{final}}(x, y) \;=\; \max\big(M_c(x, y),\, M_{\text{ecs}}(x, y)\big), \tag{4}$$

thereby covering both core and peripheral object regions and producing a more complete localization under weak supervision.

2.3 Contrastive Learning of Class-Agnostic Activation Mapping

Contrastive Class-Agnostic Activation Mapping (CCAM) [5] learns to separate foreground from background without class labels. From an input image X_i, a ResNet-50 backbone produces feature maps Z_i, and a lightweight Disentangler head generates a class-agnostic activation map $P_i \in [0, 1]^{H \times W}$. The background map is $1 - P_i$. We pool features into descriptors:

$$v_i^f = P_i\, Z_i^T, \qquad v_i^b = (1 - P_i)\, Z_i^T,$$

with $v_i^f, v_i^b \in \mathbb{R}^{1 \times C}$.

We then apply a contrastive loss that (1) pushes all foreground–background pairs (v_i^f, v_j^b) apart,

$$\mathcal{L}_{\text{NEG}} = -\frac{1}{n^2} \sum_{i,j} \log\big(1 - \text{sim}(v_i^f, v_j^b)\big),$$

and (2) pulls semantically similar pairs within foregrounds and within backgrounds together,

$$\mathcal{L}_{\text{POS}} = -\frac{1}{n(n-1)} \sum_{i \neq j} \big(w_{i,j}^f \log \text{sim}(v_i^f, v_j^f) + \; w_{i,j}^b \log \text{sim}(v_i^b, v_j^b)\big),$$

where $\text{sim}(a, b) = a^\top b / (\|a\| \|b\|)$ and $w_{i,j}^{\{\cdot\}}$ are rank-based weights. The total loss $\mathcal{L} = \mathcal{L}_{\text{NEG}} + \mathcal{L}_{\text{POS}}$ refines P_i to accurately delineate full object regions.

2.4 Open-Ended Exploration

Dataset Augmentation Incorporate background-only images as negative samples to improve the CAM related methods' performance.

Class-Specific Training Split images by bleed (i.e. Abyssinian) and fine-tune separate ResNet-50 branches to specialise occlusion handling.

These modifications remain fully weakly supervised without adding pixel-level labels.

3 Experiments

Dataset Our approach is trained and evaluated on the Oxford-IIIT Pet Dataset [3]. This dataset contains 7,349 images of 37 pet breeds (12 cat breeds and 25 dog breeds), each annotated with pixel-level segmentation masks and image-level breed labels. It is noted that only image-level labels, rather than pixel-level annotations, are employed by our methods. The original dataset is split into 80% as training images and 20% as test images. We also evaluated on the Stanford Background Dataset [1], which comprises 715 outdoor scenes from LabelMe, MSRC, PASCALVOC and Geometric Context, all annotated with semantic and geometric labels via AMT.

Setup We start with a ResNet-50 backbone pretrained on ImageNet, then fine-tune it for either 37-way pet classification or binary cat-vs-dog classification. From the resulting classifier we generate CAM, ECS-CAM, CCAM, and CCAM+ECS activation maps, which are post-processed with dense CRF and thresholded at 0.3 to produce pseudo-masks. Finally, these masks supervise the training of a U-Net segmentation model.

Table 1: Evaluation on CAM-based pseudo-mask generation

performance	**Pretrained**		**2-Class**		**37-Class**	
method	IoU	Dice	IoU	Dice	IoU	Dice
CAM	26.03	40.51	12.73	21.7	20.89	33.19
CAM+ECS	25.68	40.13	13.64	23.13	20.98	33.32
CCAM	68.21	80.32	41.07	55.85	15.1	23.93
CCAM+ECS	—	—	—	—	24.59	37.31

Evaluation metrics Two widely used metrics for semantic segmentation evaluate the overlap between the predicted mask S and the ground-truth mask G:

$$\text{IoU} = \frac{|S \cap G|}{|S \cup G|}, \quad \text{Dice} = \frac{2\,|S \cap G|}{|S| + |G|}. \tag{5}$$

Open-Ended Experiments We augment the training set with background-only images, expanding the classifier to 38-way (pets+background) or 3-way (cat, dog +background). This additional "bg" class helps suppress false positives. In addition to background augmentation, we explore a pet-specific contrastive fine-tuning stage for CCAM. We treat each of the 37 animal breeds as a separate category, sampling intra-class positives and inter-class negatives to strengthen feature discrimination. This targeted contrastive stage produces sharper, more complete activation maps for each pet type.

4 Results

Table 2: Data enhancements performance

performance	**2-Class**		**2/bg-Class**		**37-Class**		**37/bg-Class**	
method	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
CAM	12.73	21.7	29.64	45.27	20.89	33.19	20.9	33.2
CAM+ECS	13.64	23.13	29.24	44.81	20.98	33.32	21.44	33.96
CCAM	41.07	55.85	48.58	63.27	15.1	23.93	29.08	41.76
CCAM+ECS	—	—	48.65	63.44	24.59	37.31	42.8	58.36

5 Discussion

Table 1 compares CAM, CAM+ECS, CCAM and CCAM+ECS on ImageNet-pretrained, 2-class (cat vs. dog) and 37-class (breeds) backbones. The pretrained

Table 3: Result on Class-specific training process and Baseline performance

performance	**Pretrained**		**2-Class**		**37-Class**		**37-Class/U-Net**	
method	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
CCAM	68.21	80.32	52.97	67.63	59.78	73.26	59.29	73.63
CCAM+dCRF	76.02	85.64	51.15	64.99	59.79	72.32	61.58	75.3
U-Net	IoU: 68.25		Dice: 84.77		-		-	

model outperforms all trained variants, underscoring the benefit of large-scale pretraining. Adding ECS to CAM yields small but consistent gains. CCAM's contrastive learning greatly improves mask quality in the 2-class setting, though its advantage shrinks on 37 classes—likely due to overfitting to fine-grained breed details. CCAM+ECS on the 37-class model recovers lost performance by smoothing boundaries and fixing isolated errors; the 2-class model quickly diverges (erasing content), and the pretrained backbone fails to match when fine-tuned with cross-entropy due to different datasets.

Table 2 shows that adding the background as a contrastive class boosts both IoU and Dice for all models. In the 2-class case, IoU jumps from 12.73 to 29.64 and Dice from 21.70 to 45.27 (CAM). In the 37-class case, IoU rises from 15.10 to 29.08 and Dice from 23.93 to 41.76 (CCAM), reaching up to 42.80/58.36 with ECS. This confirms that an explicit background class helps localize salient regions. For class-specific experiments shown in Table 3, the 37-class model outperforms the 2-class model, as fine-grained breeds share more consistent regions than a simple cat/dog split. The ImageNet-pretrained backbone still achieves the highest scores, owing to its training on far more images (hundreds of breeds $\times$ thousands of images) versus 200 images per breed in our dataset. Although quantitative metrics for our methods remain close to the backbone, visualizations show noticeably cleaner masks[2]. Applying a CRF post-processing step to the pretrained model sharpens object boundaries—improving visual quality even if IoU/Dice drop slightly—while other methods see minimal change under CRF.

6 Conclusion

We have shown that diverse CAM-based localization techniques can be systematically combined to generate robust pseudo-masks from only image-level tags. Our unified benchmark revealed distinct failure modes of CAM, ECS-CAM, and CCAM, which we address via an ensemble fusion and lightweight CRF refinement. Introducing background-only augmentation and pet-specific contrastive fine-tuning further enhances mask completeness and reduces false positives. Our current study is constrained by the size and diversity of the pet dataset. In future work, we will scale to larger, more varied datasets.

[2] Example images are shown in Appendix B

References

1. Gould, S., Fulton, R., Koller, D.: Decomposing a scene into geometric and semantically consistent regions. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 1–8. IEEE, Kyoto (Sep 2009). https://doi.org/10.1109/ICCV.2009.5459211, http://ieeexplore.ieee.org/document/5459211/
2. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected CRFs with gaussian edge potentials (Oct 2012). https://doi.org/10.48550/arXiv.1210.5644, http://arxiv.org/abs/1210.5644, arXiv:1210.5644 [cs]
3. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)
4. Sun, K., Shi, H., Zhang, Z., Huang, Y.: ECS-net: improving weakly supervised semantic segmentation by using connections between class activation maps. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7263–7272. IEEE, Montreal, QC, Canada (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.00719, https://ieeexplore.ieee.org/document/9711489/, tLDR: This work uses relationships between CAMs to propose a novel weakly supervised method, Erased CAM Supervision Net (ECS-Net), which generates pixel-level labels by predicting segmentation results of those processed images, outperforming previous state-of-the-art methods.
5. Xie, J., Xiang, J., Chen, J., Hou, X., Zhao, X., Shen, L.: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation (Dec 2022). https://doi.org/10.48550/arXiv.2203.13505, http://arxiv.org/abs/2203.13505, arXiv:2203.13505 [cs]
6. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization (Dec 2015). https://doi.org/10.48550/arXiv.1512.04150, http://arxiv.org/abs/1512.04150, arXiv:1512.04150 [cs]

A Dense Conditional Random Fields

Dense Conditional Random Fields (Dense CRFs) are commonly employed as a post-processing step to refine coarse semantic segmentation outputs by enforcing spatial and appearance consistency across all pixel pairs. Given an image with pixels $i \in \{1, \ldots, N\}$, let x_i denote the label assigned to pixel i. A dense CRF defines the Gibbs energy of a labeling $x = \{x_i\}$ as

$$E(x) \;=\; \sum_i \psi_u(x_i) \;+\; \sum_{i<j} \psi_p(x_i, x_j), \tag{6}$$

where:

- $\psi_u(x_i)$ is the *unary potential* for pixel i, typically derived from a CNN's softmax log-probabilities.
- $\psi_p(x_i, x_j)$ is the *pairwise potential*, defined as

$$\psi_p(x_i, x_j) \;=\; \mu(x_i, x_j) \sum_{m=1}^{M} w^{(m)}\, k^{(m)}(f_i, f_j), \tag{7}$$

  with:
  - $\mu(x_i, x_j)$ the label compatibility function (often the Potts model, $\mu(a, b) = \mathbf{1}[a \neq b]$).
  - $k^{(m)}(f_i, f_j)$ Gaussian kernels on feature vectors f_i (e.g., pixel positions and colors), each weighted by $w^{(m)}$.

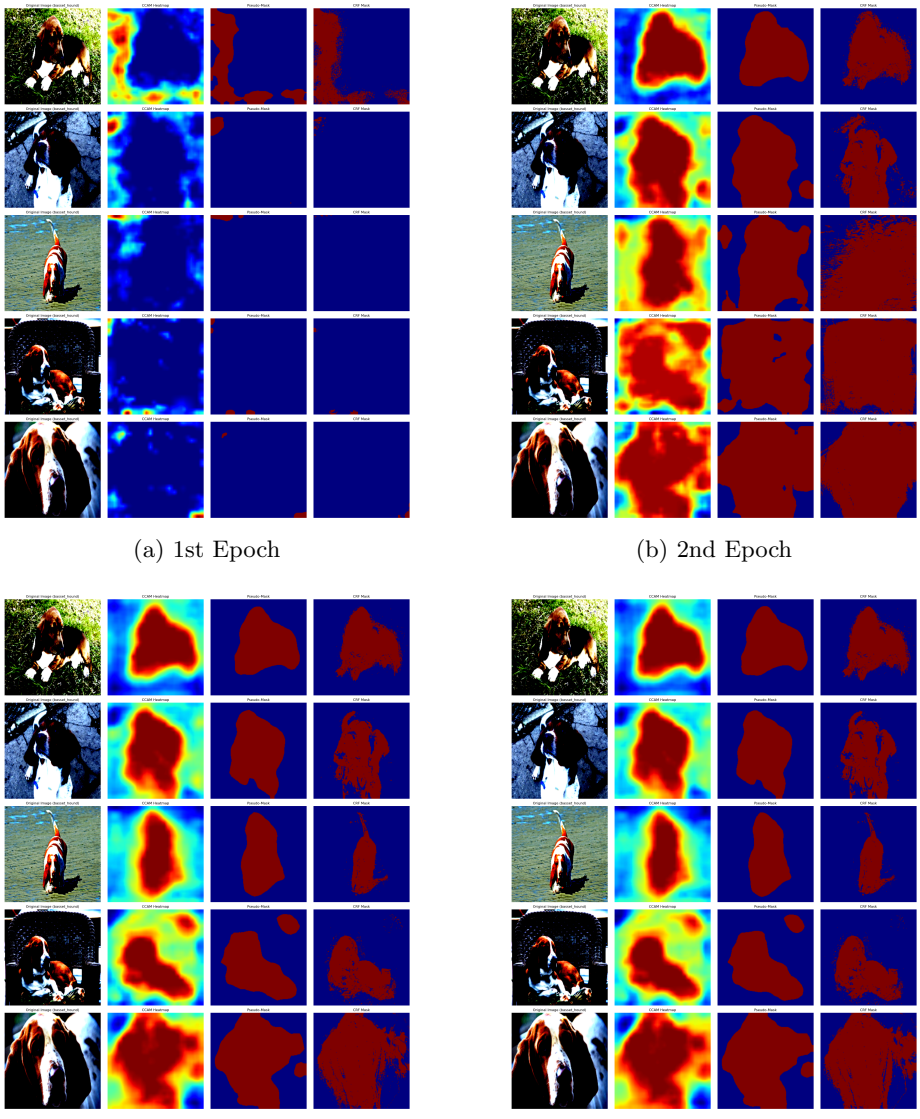
Exact inference is intractable due to the fully-connected graph; however, the mean-field approximation enables efficient approximate inference in $O(N)$ time per iteration using high-dimensional filtering (e.g. permutohedral lattice). In practice, integrating a dense CRF with CNN outputs sharpens object boundaries and removes spurious activations, yielding significant improvements in segmentation metrics with minimal computational overhead.

Algorithm 1 Mean field in fully connected CRFs

Initialize Q
while not converged **do**
  $\widetilde{Q}_i^{(m)}(l) \leftarrow \sum_{j \neq i} k^{(m)}(f_i, f_j)\, Q_j(l) \quad \forall m$ ▷ Message passing
  $\widehat{Q}_i(x_i) \leftarrow \sum_{l \in \mathcal{L}} \mu^{(m)}(x_i, l) \sum_m w^{(m)}\, \widetilde{Q}_i^{(m)}(l)$ ▷ Compatibility transform
  $Q_i(x_i) \leftarrow \exp\{-\psi_u(x_i) \;-\; \widehat{Q}_i(x_i)\}$
    normalize $Q_i(x_i)$ ▷ Local update
end while

B Example Images

Here are images on a class specific training.



(a) 1st Epoch

(b) 2nd Epoch

(c) 13th Epoch

(d) 20th Epoch

Fig. 2: Images from left to right: Original Image, CCAM, CCAM Mask and CCAM Mask + CRF