Weakly-Supervised Semantic Segmentation Using Class Activation Maps and Pseudo Masks

Yuhang Zhou

University College London
Department of Computer Science
zcabhoj@ucl.ac.uk

1 Introduction

Semantic segmentation assigns a class label to each pixel, enabling dense visual understanding. Modern models such as U-Net and DeepLabV3 achieve impressive accuracy, but require extensive pixel-level annotations, which are costly and time-consuming to acquire, especially in expert domains such as medical imaging. To reduce annotation costs, weakly supervised semantic segmentation (WSSS) uses weaker forms of supervision, such as image-level labels or bounding boxes. A common strategy is to train a classifier and extract class activation maps (CAMs) [3], which localize discriminative object regions and can be transformed into pseudo-segmentation masks for training segmentation models. In this work, we investigate how CAM-based pseudo-labels can enable effective WSSS. We evaluate the impact of thresholding strategies (fixed vs. Otsu), classifier backbones (ResNet18 vs. ResNet50), CAM variants (CAM, Grad-CAM, Grad-CAM++), and segmentation architectures (UNet vs. DeepLabV3). We also explore GrabCut post-processing to refine noisy masks and analyze trade-offs between pseudo-label quality and segmentation performance. In the Minimum Required Project (MRP), we apply this framework to the Oxford-IIIT Pet dataset using only image-level labels. A classifier generates CAMs, which are binarized and optionally refined. These masks train a segmentation model. Performance is evaluated using IoU and Dice scores, along with qualitative visualization. In the open-ended question (OEQ), we investigate whether attention maps from pre-trained Vision Transformers (ViT) can serve as pseudo-masks. We use multi-layer attention fusion and image processing (smoothing, thresholding, morphology) to generate high-quality masks without labels. We aim to evaluate whether ViT attention provides superior object coverage and robustness under weak supervision.

2 Methods

2.1 Classifier Training

We train ResNet-based image classifiers (ResNet18 and ResNet50) on the Oxford-IIIT Pet dataset using only image-level breed labels. All classifiers are initialised

with pre-trained ImageNet weights and tuned using cross-entropy loss. To improve generalisation, we apply standard data enhancements including randomly resized cropping and horizontal mirroring. For CAM extraction, we capture activations from the last convolutional block (i.e. `layer4`).

2.2 CAM-Based Pseudo Mask Generation (MRP)

We explore three CAM variants: CAM [3], Grad-CAM [2], and Grad-CAM++ [1]. Each method highlights class-discriminative regions from the classifier, which are normalized to the $[0, 1]$ range. We then apply binarization using either a fixed threshold (e.g., 0.5) or Otsu's adaptive method. To enhance mask quality, we optionally apply GrabCut, initialized from the CAM binary mask, to refine object boundaries based on color distributions and graph cuts.

2.3 Segmentation Models

We adopt two representative segmentation architectures. UNet is an encoder-decoder model with skip connections that preserves spatial detail, while DeepLabV3 uses dilated convolutions and an ASPP module on a ResNet50 backbone to capture multi-scale context. Both models are trained on the pseudo masks using binary supervision with `BCEWithLogitsLoss`, optionally combined with dice loss for class imbalance handling.

2.4 ViT Attention Pseudo Mask (OEQ)

For the open-ended question, we investigate whether attention maps from a pretrained ViT can serve as pseudo-masks without any supervision. We extract attention maps from multiple Transformer layers (e.g., 1st, 3rd, ..., 11th) and apply mean-max fusion across heads and layers. The resulting low-resolution map is upsampled to the original image size and refined through: (1) Gaussian smoothing, (2) Otsu thresholding, (3) morphological operations (e.g., dilation, erosion), and (4) largest connected component filtering. These attention-derived masks are then used to train UNet for downstream segmentation.

2.5 Fully-Supervised Baseline

As an upper bound, we train a UNet model on the pixel-level ground truth trimaps provided by the dataset. During training and evaluation, boundary pixels (label = 3) are excluded to avoid ambiguity. This serves as a fully supervised benchmark to evaluate our weakly supervised approaches.

2.6 Evaluation Metrics

All segmentation results are evaluated using two standard metrics: mean intersection over union (mIoU) and dice coefficient (F1 score). In addition to quantitative scores, we provide qualitative visualisations to analyse segmentation accuracy and mask structure.

3 Experiments

3.1 Setup

Experiments are conducted on the Oxford-IIIT Pet dataset, using the official training/testing split. ResNet classifiers are trained using image-level labels only. Pseudo-masks are generated from CAMs or ViT attention and used to train segmentation models. All evaluations are performed on the test set.

3.2 Ablation Study

We perform extensive ablation studies to analyze the impact of different design choices:

1. CAM variant: CAM vs. Grad-CAM vs. Grad-CAM++
2. Classifier backbone: ResNet18 vs. ResNet50
3. Binarization strategy: Fixed threshold (0.5) vs. Otsu thresholding
4. Post-processing: With vs. without GrabCut

3.3 Segmentation Model Comparison

To assess the robustness of segmentation architectures under weak supervision, we train both UNet and DeepLabV3 using CAM-derived pseudo-masks and compare their performance. DeepLabV3 benefits from a stronger backbone and multi-scale context modelling.

3.4 OEQ: ViT Attention Mask

We evaluate the effectiveness of ViT-generated attention masks as a fully self-supervised alternative to CAM. The masks, after processing and thresholding, are used to supervise a UNet model. We compare ViT-driven segmentation results with CAM-based methods in terms of both Dice and IoU.

4 Results

4.1 Quantitative Results

Table 4.1 shows the segmentation performance (IoU and Dice) for different combinations of pseudo-mask generation methods and segmentation architectures. A consistent observation is that DeepLabV3 outperforms UNet in almost all settings. This suggests that DeepLabV3's design - particularly its use of dilated convolutions and Atrous Spatial Pyramid Pooling (ASPP) - provides a superior ability to reconstruct spatial structure from noisy or incomplete monitoring.

Another key finding is the advantage of adaptive binarisation. Otsu thresholding consistently produces higher segmentation scores than fixed thresholds for

both ResNet18 and ResNet50 classifiers. This highlights its robustness in dealing with different CAM activation distributions, especially when CAM intensity varies across samples.

Among the CAM variants, Grad-CAM++ shows noticeable improvements over vanilla CAM, especially for ResNet50. By exploiting higher order gradients, Grad-CAM++ produces activation maps with better spatial localisation, resulting in improved pseudo-mask quality. With UNet, Grad-CAM++ achieves the best Dice score (0.613) among all non-refined CAM methods.

Most interestingly, post-processing with GrabCut yields the most significant performance gains. When applied after Otsu-thresholded CAMs, it improves mask coherence and boundary sharpness. The DeepLabV3 model trained on these refined masks achieves 0.744 IoU and 0.841 Dice, outperforming even the fully supervised UNet trained on ground-truth trimaps.

Finally, the ViT based self-supervised attention mask provides a compelling baseline. Without the need for labels or classifier training, the ViT-derived masks achieve competitive results (Dice = 0.522). This shows the potential of pre-trained transformer attention masks as an independent source of weak supervision for segmentation tasks.

Setup	**Segmentation Model**	**IoU**	**Dice**
CAM (ResNet18, threshold = 0.5)	UNet	0.427	0.583
CAM (ResNet18, threshold = Otsu)	UNet	0.434	0.589
CAM (ResNet18, threshold = 0.5)	DeepLabV3	0.509	0.661
CAM (ResNet18, threshold = Otsu)	DeepLabV3	**0.531**	**0.684**
CAM (ResNet50, threshold = 0.5)	UNet	0.249	0.380
CAM (ResNet50, threshold = Otsu)	UNet	0.419	0.577
CAM (ResNet50, threshold = 0.5)	DeepLabV3	0.329	0.487
CAM (ResNet50, threshold = Otsu)	DeepLabV3	0.412	0.576
Grad-CAM (ResNet50, threshold = 0.5)	UNet	0.436	0.596
Grad-CAM++ (ResNet50, threshold = 0.5)	UNet	0.459	0.613
CAM + GrabCut (Otsu Init)	UNet	0.477	0.622
CAM + GrabCut (Otsu Init)	DeepLabV3	**0.744**	**0.841**
ViT Attention Mask	UNet	0.375	0.522
Supervised GT	UNet	**0.635**	**0.760**

Table 1. Comparison of segmentation performance across all segmentation models

4.2 Qualitative Results

Figure 1 shows a comparison of the segmentation results produced by different CAM-based configurations. It can be seen that CAM (ResNet18) with a fixed threshold of 0.5 produces relatively coarse masks, while the use of Otsu thresholds results in improved object coverage. Switching to DeepLabV3 further improves segmentation quality due to its stronger backbone and spatial pyramid

pooling. CAM masks based on ResNet50 appear slightly noisier than ResNet18 in our experiments.

Figure 3 shows the effect of applying GrabCut post-processing to CAM masks. Boundary quality is significantly improved and background noise is effectively suppressed, resulting in masks that are much closer to ground truth quality.

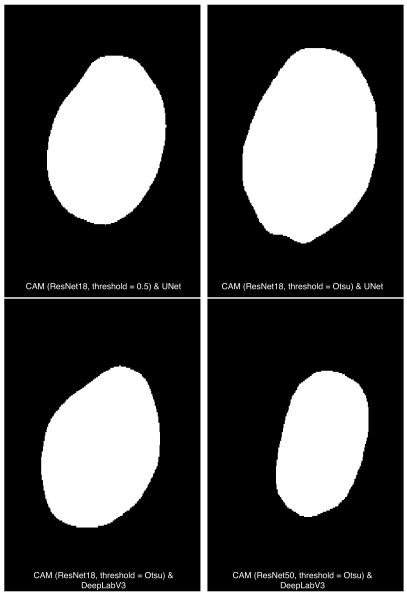

Fig. 1. Comparison of pseudo masks generated under different CAM settings. Top: UNet trained with CAM from ResNet18. Bottom: DeepLabV3 with ResNet18 and ResNet50.

Fig. 2. Example of ViT-based WSSS. Left: original image. Right: binary pseudo mask generated from ViT attention. Despite lacking pixel-level supervision, the attention-based mask captures the object structure effectively.

Fig. 3. Example of GrabCut refinement applied to a CAM-based mask. Left to right: original CAM heatmap, initial mask, and final refined mask.

Finally, Figure 2 shows the segmentation mask generated from Vision Transformer (ViT) attention maps. Despite the lack of pixel-level supervision, the attention-based pseudo-mask captures the object structure well and achieves competitive border sharpness, demonstrating the potential of ViT as an alternative weakly supervised signal.

5 Discussion

Our results highlight key characteristics of WSSS using CAM-based and transformer-based pseudo-masks. CAM-derived masks are inherently limited by localization bias - they tend to focus on the most discriminative parts of the object, resulting in incomplete segmentation. Grad-CAM++ helps to alleviate this problem by incorporating higher order gradients, resulting in more complete object coverage. Post-processing techniques such as GrabCut also play an important role by introducing low-level image priors. Although not trainable, GrabCut significantly improves boundary accuracy and spatial consistency, demonstrating the value of combining high-level activations with traditional vision heuristics.

Decoder selection is also critical under weak supervision. DeepLabV3 consistently outperforms UNet, even when trained on the same masks, suggesting that architectural components such as stronger encoders and ASPP are better at reconstructing structure from noisy labels. Furthermore, our results show that ViT attention, although self-supervised, offers a promising alternative: it achieves competitive foreground localisation without any manual annotation, highlighting the potential of transformer attention as a general-purpose weak signal.

Despite these promising results, our framework has two major limitations. First, CAM-based masks often capture only salient object parts, leading to under-segmentation. Second, our study is limited to binary segmentation; extension to multi-class settings may require class-aware activation maps or fusion strategies.

Future work could explore refinement modules such as CRFs, edge detection or saliency guidance to improve mask quality. Prompt-based ViTs (e.g. SAM, DINOv2) may also offer improved attention resolution and category awareness for multi-class weak supervision.

6 Conclusion

In this coursework, we investigated a WSSS framework using CAM-based pseudo masks. We evaluated various classifier backbones, CAM variants, thresholding methods, post-processing strategies, and segmentation architectures. Our experiments demonstrate that Grad-CAM++, Otsu binarization, and GrabCut refinement lead to significant performance gains. We further explored a novel ViT-based attention mechanism that generates effective unsupervised masks. Overall, the results validate the potential of weakly-supervised learning and attention-driven pseudo labels in semantic segmentation.

References

1. Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Improved visual explanations for deep convolutional networks. arXiv preprint arXiv:1710.11063 (2018), https://arxiv.org/abs/1710.11063, accessed Apr. 14, 2025
2. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. arXiv preprint arXiv:1610.02391 (2017), https://arxiv.org/abs/1610.02391, accessed Apr. 14, 2025
3. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. arXiv preprint arXiv:1512.04150 (2016), https://arxiv.org/abs/1512.04150, accessed Apr. 14, 2025