Efficient Weakly-Supervised Learning using Multi-Headed Segmentation for Pet Image Segmentation

Victor Fizesan

University College London

1 Introduction

Semantic segmentation is a fundamental problem in computer vision for assigning class labels to pixels in image. In the context of pet images, it enables use cases such as background removal for photo editing, augmented reality filters, and medical imaging analysis [16,19]. Traditional segmentation methods such as Fully Convolutional Networks (FCNs) [14], U-Net [16], and DeepLab [2] have achieved strong performance, but rely on large-scale pixel-level annotations, which are expensive and time-consuming.

This motivates weakly-supervised semantic segmentation (WSSS), which learns segmentation from cheaper but less informative labels such as image-level tags, bounding boxes, scribbles, or points [1,15,23]. These labels exist on a spectrum — from class labels, which are easiest to obtain but least spatially informative, to bounding boxes, which provide rough object localization at higher annotation cost. Among them, image-level class labels are the most accessible but lack explicit spatial guidance [4].

A breakthrough in WSSS came with Class Activation Maps (CAMs), which allow classification-trained CNNs to generate coarse heatmaps for regions most relevant to predicting a class [21]. These heatmaps serve as pseudo-labels, enabling segmentation without pixel-level supervision. However, CAMs tend to highlight only the most discriminative object parts — such as a pet's face or fur pattern — and miss the full spatial extent of the object [5] [7]. Improving CAM completeness has been the focus of several extensions such as SEC [10], SEAM [18], and Puzzle-CAM [7]. Other forms of weak supervision, such as bounding boxes [8] or external mask generators like the Segment Anything Model (SAM) [3,9,11,19], offer stronger spatial priors, but vary in reliability.

In this work, we focus on pet image segmentation using the Oxford-IIIT Pet dataset. We combine CAMs from image labels, bounding boxes, and SAM-generated masks within a unified U-Net-based model that uses multiple segmentation heads. Each segmentation head in our model is trained using a different weak supervision signal. Unlike traditional multi-stage pipelines that separate pseudo-label generation from segmentation training, our approach combines both into a single, unified architecture. We hypothesize that even coarse supervision signals can be highly effective when paired with the right architectural inductive biases.

Project Objectives

Minimum Required Project (MRP): We ask: *How does training with a small fully-supervised dataset and a large set of weakly-supervised pseudo-labels (e.g., CAMs) affect segmentation performance?* To evaluate this, we compare a baseline model trained only on pixel-level trimaps with models that also incorporate weak labels, focusing on low-resource settings with 5–10% annotated data.

Open-Ended Question (OEQ): Beyond the MRP, we explore: *How do different types of weak supervision — CAMs, bounding boxes, and SAM-generated masks — differ in their impact on segmentation quality?* These supervision signals vary in cost, spatial precision, and semantic informativeness. Understanding their relative strengths can help guide the design of efficient segmentation systems in settings with limited annotations.

2 Methods

In this work, we adopt a hybrid supervision framework that combines weak and strong signals within a unified multi-headed segmentation model. Each segmentation head is trained using a distinct label type: CAMs from image-level tags, binary masks from bounding boxes, SAM-generated masks, and a small subset of fully-supervised trimaps. This design enables both comparative and complementary learning from diverse supervision sources. The approach draws inspiration from semi-supervised learning across domains, including consistency training in object detection [6], self-supervised methods in image classification [20], and natural language processing [13].

2.1 Architecture

Our model uses a U-Net architecture [16] with a ResNet34 encoder pretrained on ImageNet. The encoder compresses spatial information through a series of downsampling convolutional blocks, while the decoder restores resolution through upsampling and skip connections. These skip connections bridge low-level spatial detail with high-level semantics, allowing precise localization in the output mask. This shared backbone feeds into four parallel segmentation heads (see Section 2.5), each supervised by a different signal.

2.2 Class Activation Map Generation

Class Activation Maps (CAMs) identify spatial regions most relevant to a classification decision. We train a ResNet-50 multi-class image classifier, and extract CAMs by computing a weighted sum over the final convolutional feature maps. These heatmaps are then upsampled to match the original image resolution, producing coarse spatial maps that highlight discriminative regions [22,17].

In our architecture, CAMs are treated as soft pseudo-labels and are used to supervise a dedicated segmentation head. This makes CAMs an efficient, fully scalable form of weak supervision applicable to every training image. However, their limited spatial extent motivates inclusion of additional supervision signals.

2.3 Bounding Box Processing

Bounding boxes which locate the pet head are available for half of the images. They are converted to binary masks used to supervise a separate model head. Though spatially coarse, these labels offer object-localization priors with minimal annotation cost, as seen in prior work such as BB-UNet [8].

2.4 Segment Anything Model (SAM)

SAM is a foundation model trained on over a billion masks with a promptable architecture that supports inputs in the form of points, boxes, or texts [9]. It generalizes to unseen domains and provides high-quality object boundaries [3,12].

We use SAM to generate segmentation masks for our dataset using bounding-box points as an argument to this model. These SAM masks, used as supervision for a dedicated segmentation head, offer refined spatial supervision and have been shown to outperform conventional weak labels in recent literature [11,19].

2.5 Multi-Headed Segmentation Architecture

Our proposed architecture combines the above weak supervision signals using a multi-headed design. A shared ResNet34-based U-Net backbone extracts image features, which are then processed by four parallel segmentation heads, each supervised by a distinct label type:

- **Trimap Head:** Supervised on a small subset of fully-annotated pixel-level trimaps (e.g., 10% of the training set).
- **CAM Head:** Supervised using heatmaps from image-level class labels.
- **Bounding Box Head:** Trained on binary bounding box masks
- **SAM Head:** Uses masks generated by the Segment Anything Model.

Each head consists of lightweight convolutional layers that map shared features into binary segmentation masks. During training, all heads are optimized jointly using a weighted sum of segmentation losses. In the final 20% of training, we backpropagate gradients only from the Trimap head. This late-stage supervision shift allows the model to consolidate knowledge from weak labels into the primary task of fine-grained segmentation.

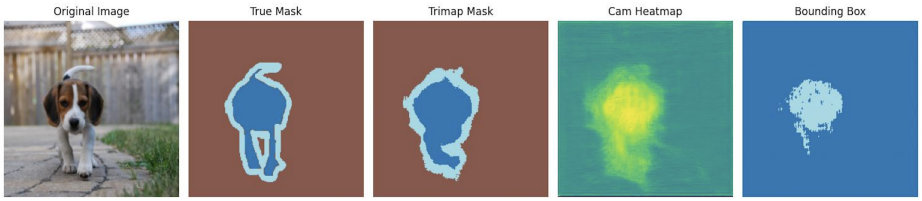

Fig. 1. Image segmentation maps from different supervision sources

3 Experiments

3.1 Dataset and Preprocessing

We evaluate weak supervision strategies using the Oxford-IIIT Pet dataset, which contains 7,000 images of 37 breeds, each annotated with a class label, a head bounding box (available for 50%), and a pixel-level trimap mask.

Images are resized to 256×256 and normalized using ImageNet statistics. The dataset is split 80/20 for training and validation. To simulate low-resource settings, we vary trimap coverage in the training set (2.5%, 5%, 10%, 20%, 33%, 50%), with remaining images either excluded (baseline) or supervised using CAMs, bounding boxes, or SAM masks.

3.2 Supervision Configurations

Each training run uses one of the following supervision configurations:

- **Baseline:** Standard U-Net trained only on the available trimap annotations. No weak supervision is used.
- **CAM Supervision:** Model trained using soft pseudo-labels from Class Activation Maps (CAMs) generated via a ResNet-50 classifier.
- **Bounding Box Supervision:** Model trained using binary masks derived from head bounding boxes.
- **SAM Supervision:** Model trained on high-quality masks generated by the Segment Anything Model (SAM) using bounding box prompts.
- **Fully Supervised:** A reference model trained only on trimaps with no weak supervision, included for baseline comparison.

3.3 Training Setup

All models are trained for 50 epochs using an Adam optimizer with an initial learning rate of 1×10^{-4}, scheduled by a Reduce-on-Plateau strategy that halves the learning rate after stagnating validation loss. The encoder is initialized with ImageNet-pretrained weights and fine-tuned throughout training.

Supervision signals are incorporated using tailored loss functions: pixel-wise cross-entropy for trimap and bounding box masks, and a sigmoid-based loss for CAMs to align predictions with soft heatmaps. Losses from all heads are combined using weighted sums. In the final 20% of epochs, only the Trimap loss is backpropagated to focus training on the fully-supervised objective.

3.4 Evaluation Metrics

Model performance is evaluated on the held-out validation set using the following metrics:

- **Pixel Accuracy:** The percentage of correctly classified pixels.
- **Recall:** The proportion of true foreground pixels correctly identified.
- **Jaccard Index:** Intersection-over-union in predicted and ground-truth masks.
- **Dice Score:** The harmonic mean of precision and recall.

We report Dice scores across trimap coverage levels for all supervision types (Table 1), and compare weak vs. fully supervised setups at 2.5% trimap coverage (Table 2). Additionally, we qualitatively inspect segmentation outputs from each head to understand differences in spatial quality and failure modes.

4 Results

Table 1. Dice Scores across weakly supervised configurations at varying data proportions.

Model	2.5%	5%	10%	20%	33%	50%
Baseline (B)	0.7208	0.8205	0.8585	0.8749	0.8798	0.8107
B + CAM	0.8051	0.8388	0.8587	0.8739	0.8789	0.8047
B + SAM	0.8140	0.8467	0.8679	0.8784	0.8820	0.8331
B + BBox	0.7855	0.8395	0.8637	0.8699	0.8779	0.8069
Fully Supervised	0.7430	0.8155	0.8687	0.8753	0.8806	0.8176

Table 2. Comparison at 2.5% supervision: the weakly-supervised model outperforms full supervision across all metrics, highlighting the effectiveness of SAM-based labels in low-data settings.

Model	Accuracy	Recall	IoU	Dice
Fully Supervised	0.8183	0.7696	0.5902	0.7208
Weakly Supervised	0.8190	0.8289	0.6112	0.7430

The results show that weak supervision is most effective in the lowest-data regime. At just 2.5% supervision, weakly-supervised models outperform the fully supervised counterpart across all metrics (Table 2), with a +2.2 point improvement in Dice score. This suggests that high-quality pseudo-labels like those from SAM can meaningfully compensate for limited annotations, even outperforming models trained solely on scarce ground-truth data.

Among individual signals, SAM consistently yields the strongest gains across all supervision levels. At 2.5% and 5%, SAM boosts Dice scores over the baseline by +9 and +2.6 points, respectively (Table 1). This confirms SAM's effectiveness as a strong spatial prior and highlights its value in early-stage supervision.

CAMs and bounding boxes also help in low-resource settings, though to a lesser degree. CAMs offer the largest relative gain at 2.5%, likely due to their scalability across the full dataset. However, their contribution quickly plateaus, and in higher supervision settings (e.g., 50%), both CAM and BBox models underperform the baseline. This suggests that these signals may introduce noise or overly coarse priors once more precise trimaps become available.

Interestingly, the fully supervised model lags behind the weakly-supervised variants at 2.5%, which may reflect overfitting to a small and possibly non-representative subset of trimaps. Meanwhile, the baseline model trained only on sparse trimaps performs surprisingly well at 10–33%, likely due to architectural priors and the dataset's regular structure. At 50%, performance drops slightly across all models, possibly due to overfitting or noisy labels in the larger supervision set.

Overall, these results show that SAM-based weak supervision is highly effective at low annotation levels, while weaker or noisier signals like CAMs and bounding boxes should be used selectively, especially as supervision scales up.

5 Discussion

Our results show that high-quality weak supervision—especially from SAM—can outperform fully-supervised training in extreme low-data regimes. At 2.5% trimap coverage, weakly-supervised models achieved better Dice and IoU scores than their fully-supervised counterparts, highlighting the strength of structured pseudo-labels when ground truth is scarce. However, weaker signals like CAMs and bounding boxes showed inconsistent benefit, with diminishing returns or even slight regressions at higher supervision levels.

Notably, performance dipped slightly at 50% supervision across all methods. This may indicate overfitting or noise in the annotated masks, suggesting that more data does not always translate to better generalization without regularization.

A major limitation of this work is that all models were trained with a single seed. Given the known sensitivity of segmentation networks to initialization and data order, multiple trials would be needed to establish statistical robustness.

Future work could incorporate seed averaging or model ensembling to reduce variance, and experiment with adaptive loss weighting to better balance signal quality. Extending evaluation to more complex or cluttered datasets would also test the generalizability of these findings beyond the structured nature of Oxford-IIIT Pet.

6 Conclusion

This work explored weakly-supervised semantic segmentation under limited annotation budgets, using CAMs, bounding boxes, and SAM-generated masks. We introduced a multi-headed U-Net architecture to integrate these signals and evaluated it on the Oxford-IIIT Pet dataset across varying levels of trimap supervision.

In addressing the MRP, we showed that supplementing a small set of ground-truth labels with weak supervision—especially from SAM—significantly improves segmentation quality in low-resource regimes. For the OEQ, we found that not all weak labels are equally effective: while SAM provided consistent gains, CAMs and bounding boxes offered limited or inconsistent benefit, particularly at higher supervision levels.

Overall, our findings reinforce that weak supervision can serve as a powerful substitute for dense annotation, but its effectiveness hinges on the quality and integration of the signals used. Future systems should prioritize high-quality pseudo-labels and treat supervision types selectively, rather than uniformly, to maximize performance under constrained labeling budgets.

References

1. Chan, L., Hosseini, M.S., Plataniotis, K.N.: A comprehensive analysis of weakly-supervised semantic segmentation in different image domains. International Journal of Computer Vision **129**(2), 361–384 (Sep 2020). https://doi.org/10.1007/s11263-020-01373-4, http://dx.doi.org/10.1007/s11263-020-01373-4
2. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation (2017), https://arxiv.org/abs/1706.05587
3. Chen, T., Mai, Z., Li, R., lun Chao, W.: Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation (2023), https://arxiv.org/abs/2305.05803
4. Chen, Z., Sun, Q.: Weakly-supervised semantic segmentation with image-level labels: from traditional models to foundation models (2024), https://arxiv.org/abs/2310.13026
5. Chen, Z., Wang, T., Wu, X., Hua, X.S., Zhang, H., Sun, Q.: Class re-activation maps for weakly-supervised semantic segmentation (2022), https://arxiv.org/abs/2203.00962
6. Jeong, J., Lee, S., Kim, J., Kwak, N.: Consistency-based semi-supervised learning for object detection. Advances in neural information processing systems **32** (2019)
7. Jo, S., Yu, I.J.: Puzzle-cam: Improved localization via matching partial and full features. In: 2021 IEEE International Conference on Image Processing (ICIP). p. 639–643. IEEE (Sep 2021). https://doi.org/10.1109/icip42928.2021.9506058, http://dx.doi.org/10.1109/ICIP42928.2021.9506058
8. el Jurdi, R., Petitjean, C., Honeine, P., Abdallah, F.: Bb-unet: U-net with bounding box prior. IEEE Journal of Selected Topics in Signal Processing **PP**, 1–1 (06 2020). https://doi.org/10.1109/JSTSP.2020.3001502
9. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
10. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 695–711. Springer International Publishing, Cham (2016)
11. Kweon, H., Yoon, K.J.: From sam to cams: Exploring segment anything model for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19499–19509 (June 2024)
12. Kweon, H., Yoon, K.J.: From sam to cams: Exploring segment anything model for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19499–19509 (2024)
13. Liang, P.: Semi-supervised learning for natural language. Ph.D. thesis, Massachusetts Institute of Technology (2005)
14. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation (2015), https://arxiv.org/abs/1411.4038
15. Lu, Z., Chen, D., Xue, D.: Survey of weakly supervised semantic segmentation methods. In: 2018 Chinese Control And Decision Conference (CCDC). pp. 1176–1180 (2018). https://doi.org/10.1109/CCDC.2018.8407307
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015), https://arxiv.org/abs/1505.04597

17. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806 (2014)
18. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12272–12281 (2020). https://doi.org/10.1109/CVPR42600.2020.01229
19. Yue, X., Liu, X., Zhao, Q., Li, J., Song, C., Liu, S., Yang, Z., Fu, G.: Morphology-enhanced cam-guided sam for weakly supervised breast lesion segmentation (2024), https://arxiv.org/abs/2311.11176
20. Zhai, X., Oliver, A., Kolesnikov, A., Beyer, L.: S4l: Self-supervised semi-supervised learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1476–1485 (2019)
21. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization (2015), https://arxiv.org/abs/1512.04150
22. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)
23. Zhu, K., Xiong, N.N., Lu, M.: A survey of weakly-supervised semantic segmentation. In: 2023 IEEE 9th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS). pp. 10–15 (2023). https://doi.org/10.1109/BigDataSecurity-HPSC-IDS58521.2023.00013