Weakly-Supervised Pet Segmentation with ViTs: FineCAM, Decoders, and Supervision Trade-offs

Hindy Rossignol[1]

University College London, London, UK
hindy.rossignol.21@ucl.ac.uk

1 Introduction

Motivation. Semantic segmentation is the process of assigning a meaningful object category (like "cat", "dog", or "background") to each individual pixel in an image. It plays a critical role in computer vision applications such as medical imaging and robotics. Traditionally, training models for such tasks is performed using fully-supervised approaches, which require large-scale pixel-level annotations. However, such data is costly and time-consuming to produce due to the need for extensive human labor.

Background. Weakly-supervised semantic segmentation (WSSS) addresses these limitations by training models using only weak labels such as image-level tags [1]. Early WSSS approaches often leveraged Class Activation Maps (CAMs) from convolutional neural networks (CNNs) to extract spatial cues from classification-trained networks, but these techniques struggle to precisely localize object regions [1].

Literature Review. Recent advances in attention mechanisms and self-supervised learning have significantly improved WSSS capabilities. Self-supervised attention methods such as SEAM [7] have improved spatial consistency in weakly-supervised semantic segmentation, helping refine object localization beyond the limits of traditional CAMs. At the same time, Vision Transformers (ViTs) have emerged as powerful alternatives to CNNs, offering global contextual representations and richer semantic features [3]. Building on recent advances, the WeakTr framework [2] introduces a ViT-based WSSS pipeline that uses Adaptive Attention Fusion (AAF) to refine CAMs into FineCAMs by aggregating attention across transformer heads and layers. This approach helps address two common problems with traditional CAM-based methods: they often highlight only small, high-confidence regions of the object (spatial sparsity), and their boundaries are imprecise (coarse). We address two core research questions:

Minimum Required Project (MRP). Which architectural and hyperparameter choices (e.g., ViT backbone type, decoder complexity, FineCAM threshold) yield the best segmentation performance under fully weak supervision? To address this, we systematically evaluate combinations of backbones, decoders, and thresholds on the Oxford-IIIT Pet Dataset, using metrics such as IoU, Dice, and pixel accuracy.

Open-Ended Question (OEQ). How does the segmentation performance change as we incrementally introduce a small proportion of ground truth masks (10–40%)? This part assesses the trade-off between annotation cost and model accuracy, helping identify the most cost-effective supervision level.

2 Methods

2.1 Weakly Supervised Semantic Segmentation Pipeline (MRP)

Our WSSS comprises five parts:

1. **ViT Classifier Training**: A Vision Transformer (ViT) backbone is fine-tuned to classify images from the Oxford-IIIT Pet Dataset. The classification head is replaced with a linear layer matching the number of classes. We use only image-level labels during this stage.
2. **CAM Generation and FineCAM Refinement**: After training, the ViT produces Class Activation Maps (CAMs) by tracing gradients back to the patch tokens. These CAMs are refined into FineCAMs using Adaptive Attention Fusion (AAF), which aggregates attention weights across multiple heads and layers to enhance object localization. Unlike the original WeakTr paper, we do not use Conditional Random Fields (CRF) for post-processing.
3. **CNN Decoder Training with Gradient Clipping**: We train a lightweight CNN decoder to reconstruct the FineCAMs from ViT patch tokens. During this stage, we apply pixel-weighted loss masking to suppress noisy activations and stabilize training. This decoder is trained separately from the ViT encoder, simplifying optimization and improving modularity.
4. **Sigmoid Binarization and Pseudo-mask Generation**: The decoder's outputs are passed through a sigmoid function and then compared to the FineCAMs using pixel-wise Binary Cross Entropy (BCE) loss. We compute a loss map, where each pixel's value reflects how poorly it was reconstructed. Low-loss (well-predicted) pixels are considered reliable and retained in the pseudo-mask, while uncertain pixels are masked out. This method provides a more nuanced binarization than raw thresholding, as it explicitly uses model confidence.
5. **Final Segmentation Training and Comparison**: The generated pseudo-masks serve as supervision for training a segmentation model (LRASPP-MobileNetV3) [5]. This is our WSSS baseline, which we compare against the fully-supervised version of the same model using ground-truth masks.

2.2 Hybrid Supervision (OEQ)

To assess the impact of partial strong supervision, we experiment with a hybrid training setup that combines the pseudo-masks generated by our decoder with varying fractions of ground truth masks. A proportion $x \in \{0.0, 0.1, 0.2, 0.3, 0.4\}$ of samples use the ground truth masks, while the remaining $1 - x$ rely on the pseudo-masks. This approach allows us to evaluate the trade-off between annotation cost and segmentation quality, and to analyze whether sparse supervision improves performance and narrows the gap with fully supervised models.

3 Experiments

3.1 Ablation Study: Architectural Design and Threshold Variants

Hypothesis 1: Certain architectural configurations yield significantly better segmentation performance under WeakTr supervision.

To identify the most effective configuration of our pipeline, we conduct an ablation study by varying three key components of the MRP pipeline:

- **ViT Backbone:** We evaluate multiple transformer sizes including ViT-Small and DeiT-Tiny for comparison (see Table 1).
- **Decoder Architecture:** We experiment with lightweight CNN decoders of increasing complexity (Small, Medium, Large) to analyze the trade-off between segmentation performance and model size (see Table 1).
- **FineCAM Thresholding:** We apply several binarization thresholds on the FineCAM heatmaps before using them to train the decoder. This affects the sharpness and confidence of pseudo masks.

All possible combinations of these three components are being systematically trained and then, evaluated on the test set using Intersection over Union (IoU), DICE, and pixel-wise accuracy. These results allow us to compare the performance impacts of each architectural and thresholding choice.

Fully Supervised Baseline: To compare our weakly supervised results, we include a fully supervised segmentation model (LRASPP-MobileNetV3) trained using all available ground truth masks in the pet dataset.

Table 1. ViT Backbone and Decoder Configurations

Vision Transformers					**CNN Decoders (Input:** $[C, 14, 14]$**)**		
Model	**Type**	**#Layers**	**Hidden Dim** (C)	**#Heads**	**Decoder**	**Channels**	**Output Size**
DeiT-Tiny	DeiT	12	192	3	Small	C→64→32→16→1	$[1, 224, 224]$
ViT-Small	ViT	12	384	6	Medium	C→128→64→32→1	$[1, 224, 224]$
					Large	C→256→128→64→1	$[1, 224, 224]$

3.2 Further Experiments: Hybrid Supervision

Hypothesis 2: Introducing a small proportion of ground truth masks (10–40%) leads to non-linear improvements in segmentation quality, revealing a trade-off between annotation effort and performance (sweet spots).

After electing the best architectures (ViT-Small and DeiT-Tiny), we apply the methods evoked in section 2.2 to quantify the impact of gradually increasing the proportion of real masks. Thanks to this, we determine the best possible threshold, by model size.

4 Results

4.1 Ablation Study

Table 2. WSSS Performance for various ViT and decoder configurations. The best-performing binarization threshold θ is used for each. The Fully Supervised Model used for comparison is LRASPP-MobileNetV3.

ViT Model	Decoder Size	θ	IoU	Dice	Accuracy	Precision	Recall
`DeiT-Tiny`	Small	0.25	**75.00**	**85.71**	**79.35**	83.50	**88.05**
	Medium	0.25	74.00	85.06	78.99	**85.10**	85.02
	Large	0.25	72.42	84.01	78.45	87.88	80.46
`ViT-Small`	Small	0.15	**75.26**	**85.89**	**77.84**	**77.82**	95.82
	Medium	0.15	75.06	85.75	77.37	76.95	96.83
	Large	0.15	75.00	85.71	77.17	76.55	**97.37**
Fully Supervised Model			**93.16**	**96.46**	**95.05**	**97.16**	**95.80**

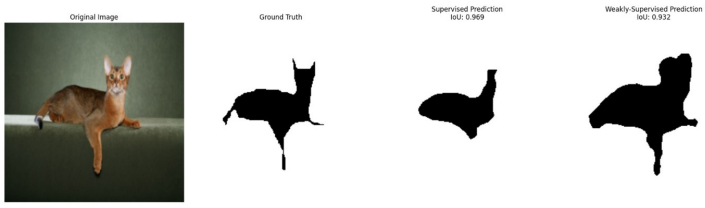

Fig. 1. Example of pseudo-mask generated by our ViT-Small vs. ground truth and fully-supervised masks. Our WSSS prediction captures subtle object parts like the cat's arms, missed by the supervised model.

Table 2 shows that both DeiT-Tiny and ViT-Small achieve promising results under weak supervision. ViT-Small with a small decoder and $\theta = 0.15$ offers the strongest performance (75.26 IoU, 85.89 Dice). This best configuration is only 10.57% lower than the fully-supervised Dice score (96.46), without using any pixel-level annotations. This highlights the strong localization capabilities of ViT encoders and affirms that FineCAM-based pseudo-masks can be effective segmentation cues.

Increasing decoder size does not consistently improve performance, likely due to overfitting or oversmoothing on noisy pseudo-masks. Meanwhile, our modular setup—training the decoder separately and omitting CRF post-processing—results in a lighter pipeline. Despite ViT-Small's larger backbone (22M parameters) only displays marginal performance gains over the DeiT-Tiny (5M parameters) model (75.26% IoU versus 75.00% IoU). This suggests that DeiT-Tiny contains enough parametric capacity to create our pseudo-masks for our down stream segmentation task.

4.2 Hybrid Supervision

Table 3. WSSS performance by proportion of Ground Truth (GT) masks used in training, using a Small decoder for each ViT backbone and the best threshold.

ViT Model	Decoder	GT %	IoU	Dice	Accuracy	Precision	Recall
`DeiT-Tiny`	Small	10	78.29	87.83	82.45	85.78	89.97
		20	78.55	87.99	82.64	58.72	90.38
		30	82.08	90.16	85.88	88.49	91.89
		40	85.47	92.17	88.88	91.35	93.00
`ViT-Small`	Small	10	77.50	87.32	80.41	80.14	95.92
		20	77.50	87.32	80.40	80.14	95.92
		30	80.72	89.33	83.85	83.41	96.15
		40	84.66	91.69	87.75	87.68	96.09
Fully Supervised Model			**93.16**	**96.46**	**95.05**	**97.16**	**95.80**

Table 3 confirms that introducing even 10% of pixel-level masks improves performance beyond the fully weakly-supervised baseline. As the proportion of ground truth increases to 40%, IoU and Dice scores consistently rise—reaching 92.17 Dice for DeiT-Tiny and 91.69 for ViT-Small. Interestingly, DeiT-Tiny benefits more from partial labels, suggesting that smaller backbones create potentially better masks for training when paired with human annotated data.

5 Discussion

Our experiments demonstrate that integrating Vision Transformers with FineCAM and decoder-based pseudo-masking achieves robust segmentation performance under weak supervision. Notably, we attain a Dice score of 92.17% using only 40% ground truth masks, and 85.71% without any real masks (DeiT-Tiny + Decoder), challenging the necessity for full supervision and highlighting the potential of hybrid training methods. The modular design of our approach—featuring separate decoder training and eliminating reliance on CRFs—offers computational advantages over methods like WeakTr. This structure facilitates efficient experimentation with various architectures during development.

However, there are limitations. Our pipeline heavily depends on the classification capabilities of large pretrained ViT models, which may be restrictive for applications with limited computational resources seeking standalone solutions. Additionally, performance is sensitive to the FineCAM binarization threshold (θ). Future work could explore adaptive thresholding techniques, such as those proposed by Lee et al. [4], or alternatively, inject human assessment in the loop to discriminate FineCAMs and enhance the robustness of the pseudo-masks.

6 Conclusion

Our research demonstrates the effectiveness of Vision Transformers and FineCAM for weakly-supervised semantic segmentation. Our modular pipeline, which trains a decoder independently of the ViT backbone, achieves strong results without pixel-level supervision, reaching up to 85.89% Dice. Incorporating as little as 10–40% of ground truth masks further improves accuracy, validating hybrid supervision as a cost-efficient alternative. While threshold sensitivity and reliance on pretrained models remain challenges, our findings set a solid foundation for scalable, accessible, and lightweight segmentation.

References

1. 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 (CVPR)* (2016)
   `https://doi.org/10.48550/arXiv.1512.04150`
2. Zhu, L., Li, Y., Fang, J., Liu, Y., Xin, H., Liu, W., Wang, X.:
   WeakTr: Exploring Plain Vision Transformer for Weakly-supervised Semantic Segmentation.
   In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023)
   `https://doi.org/10.48550/arXiv.2304.01184`
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.:
   An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.
   In: *International Conference on Learning Representations (ICLR)* (2021)
   `https://doi.org/10.48550/arXiv.2010.11929`
4. Lee, D., Choi, Y., Kim, D.:
   Threshold Matters in WSSS: Adaptive Thresholding for Reliable Pixel Supervision.
   In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022)
   `https://doi.org/10.48550/arXiv.2203.16045`
5. Howard, A., Sandler, M., Chu, G., et al.:
   Searching for MobileNetV3.
   In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2019)
   `https://doi.org/10.48550/arXiv.1905.02244`
6. Parkhi, O. M., Vedaldi, A., Zisserman, A., Jawahar, C. V.:
   Cats and Dogs.
   In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2012)
   `https://www.robots.ox.ac.uk/~vgg/data/pets/`
7. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.:
   SEAM: Self-supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation.
   In: *European Conference on Computer Vision (ECCV)* (2020)
   `https://doi.org/10.48550/arXiv.2004.04581`