Improving CAM-Based Pseudo-Segmentation with Iterative Self-Training and Superpixels

Anonymous Submission

No Institute Given

1 Introduction

Semantic segmentation is a core computer vision task involving pixel-level classification. Fully supervised methods perform well but require costly, time-consuming annotations. Weakly supervised semantic segmentation (WSSS) offers an alternative using coarse labels like image-level tags, bounding boxes, or noisy masks.

Recent advances in WSSS utilise class activation maps (CAMS) to highlight discriminative regions of objects, leveraging image-level labels. Approaches such as Grad-CAM++ [9] offer class-specific localisation cues, which then can be transformed into pseudo-labels for training segmentation models. Other methods, such as PseudoSeg [11], employ iterative self-learning to refine predictions over time. Another line of work, such as Pixel to Prototype Contrast (PPC) [4], explores contrastive learning using prototype representations to align pixel embeddings with their respective class semantics. These techniques reflect the broader landscape of weak supervision described by Zhou [1], which categorises it into incomplete, inexact, and inaccurate supervision.

This project compares weakly supervised segmentation strategies using only image-level labels. The Main Research Project (MRP) investigates three WSSS methods: (1) a prototype-based contrastive model (PPC), (2) a Grad-CAM++ based pseudo-labelling approach with U-Net, and (3) a DeepLabV3 model trained on Grad-CAM++ masks. These are evaluated against a fully supervised U-Net baseline trained with dense annotations.

The Open Ended Question (OEQ) component extends the third model using a self-training pipeline. DeepLabV3 is further refined using its predictions and SLIC superpixels [8] in an iterative self-training pipeline to enhance mask quality without requiring additional supervision.

Research questions: MRP: To what extent can weak supervision based on CAMs and prototype-based learning approximate the performance of fully supervised semantic segmentation models? Which weakly supervised strategy performs best under a common training and evaluation pipeline? OEQ: Can iterative self-learning with superpixel refinement significantly enhance the quality of CAM-based pseudo labels in weakly supervised segmentation? How does the quality of pseudo-label refinement compare to architectural or training variations?

This project aims to find out the effectiveness of weakly supervised learning pipelines in scenarios where no annotations are available.

2 Methodology:

2.1 Supervised Baseline: U-Net

A fully supervised U-Net model was implemented as a performance baseline. Using PyTorch, a U-Net model was trained with pixel-level annotations from the Oxford-III Pet dataset. The network architecture follows the classical U-Net encoder-decoder structure with skip connections. Each encoder block consists of two convolutional layers followed by batch normalisation, ReLU activation, and max pooling. The decoder has the same structure as transposed convolutions and feature concatenation.

The final 1x1 convolution maps the decoder output to segmentation logits. In practice, segmentation is evaluated using a binary mask (foreground vs background). The model was trained using CrossEntropyLoss and optimised with Adam (learning rate = 0.01). The input images were resized to 256x256, normalised, and augmented with random horizontal flips and rotations. This preprocessing was shared across all models. Performance was measured using mean Intersection over Union (mIoU), typically computed on binary foreground-background segmentation. The U-Net architecture was originally proposed for biomedical image segmentation [2] and has since become a standard baseline for semantic segmentation.

2.2 Weakly Supervised Model 1: Pixel to Prototype Contrast

This model is based on the paper [4], which implements pixel-to-prototype contrastive learning using image-level labels. The model utilises a pre-trained ResNet-101 backbone to extract deep features from input images, and the ResNet-101 backbone is initialised with IMAGENET1K_V2 pretrained weights for better feature representations. A 1x1 convolutional layer projects these features to generate Class Activation Maps (CAMS), which are used to identify class-specific regions.

For computing class prototypes, the first step is to extract the top-K activated pixels per class from the CAMS. These feature vectors are L2 normalised and averaged to obtain a prototype for each class. Pixel embeddings are projected and compared to prototypes using a temperature-scaled contrastive loss. The loss maximises the similarity between a pixel and the correct class prototype while minimising similarity to other class prototypes.

SEAM [13] is used to generate initial CAMs, which guide pixel sampling for prototype contrast in PPC. CAMs guide prototype construction but are not directly evaluated. Segmentation outputs are evaluated against ground truth using mean Intersection over Union (mIoU). It should be noted that these outputs are binarised to predict foreground vs. background masks, aligned with the simplified segmentation task.

2.3 Weakly Supervised Model2: Grad-CAM++ based Pseudo labelling and Segmentation

The second model for weakly supervised learning is a two-stage training strategy, based on Grad-CAM++ [9]. The first step was to train a ResNet18 (with randomly initialised weights) image classifier using only the image labels from the dataset. Then, Grad-CAM++ is applied to generate class-discriminative activation maps from the trained classifier.

To improve localisation quality, CAMS are generated across multiple input scales (e.g., 64, 112, 224, 320, 448) and averaged. These CAMs are then thresholded using a three-region strategy: high-activation areas are labelled as foreground, low-activation areas as background, and uncertain regions are ignored. The resulting foreground and background pseudo-labels are used to supervise a U-Net model that outputs binary segmentation masks (foreground vs. background). Training uses CrossEntropyLoss (ignore index = 1) to skip uncertain pixels. The Trained U-Net is evaluated using the mean Intersection over Union (mIoU) between the model's predictions and the ground truth segmentation maps from the test set.

2.4 Weakly Supervised Model 3: Iterative Self-Learning with DeepLabV3 and Superpixels(OEQ)

This model initially uses standard weak supervision, trained on Grad-CAM++ pseudo-labels from a ResNet18 classifier. Then, DeepLabV3 with a ResNet-50 is used as a segmentation architecture. DeepLabV3 is a state-of-the-art semantic segmentation model that combines atrous spatial pyramid pooling with an encoder-decoder design for refined boundary segmentation [10]. Also, the ResNet50 backbone used in DeepLabV3 is randomly initialised and not pre-trained on ImageNet. The initial training is performed using CAM-based pseudo-labels with the CrossEntropyLoss objective, resulting in a weakly supervised training set that utilises no dense annotations.

For the OEQ part, this model is expanded into an iterative self-learning framework. After the first round of training, the model's predictions on the training data are used as the pseudo-labels. These masks are binarised to distinguish the pet foreground from the background during training. Then, a further refinement using SLIC superpixels [8] is performed to assign the dominant class label to all pixels within each superpixel. This leads to enhancing spatial coherence and segmentation consistency. Lower SLIC compactness (e.g., 0.01, 0.001) improves boundary alignment by prioritising colour similarity.

The training cycle is repeated using these refined pseudo-labels. Retraining continues iteratively, with mIoU used for evaluation. This process enables the model to enhance its supervision through its outputs, demonstrating a self-training approach under weak supervision. Our method is inspired by recent advances in pseudo-label-based weak supervision, such as PseudoSeg [11] and CSST [12]. While PseudoSeg focuses on fusing soft activation maps for better label quality, CSST demonstrates the scalability of hard pseudo-label-based iterative self-training, which aligns more closely with our proposed pipeline.

3 Experiments

3.1 MRP Comparison: Fully Supervised vs Weakly Supervised Models

To establish a performance benchmark, both fully supervised and weakly supervised models are implemented. The fully supervised U-Net model, trained on pixel-level segmentation masks from the Oxford-IIIT Pet dataset, is designed to serve as the upper-bound baseline, demonstrating what is achievable when full supervision is available. Then, the three weak models—PPC, Grad-CAM++, U-Net, and DeepLabV3 (+self-learning), are compared to the supervised model. To isolate architectural effects, DeepLabV3 was trained on the same Grad-CAM++ pseudo-labels as U-Net. This isolates the architectural effect under identical supervision. All models were evaluated using the same dataset splits, training procedures, and the mean Intersection over Union (mIoU) metric, ensuring a fair and consistent comparison.

3.2 Ablation Experiments

A set of controlled experiments was conducted to understand the impact of individual design decisions on the performance of DeepLabV3. The first ablation varied the number of training epochs (e.g., 5 vs 10) to evaluate how segmentation performance scales with longer training. Data augmentations were removed to study their effect on generalisation and robustness. Another test was conducted by setting the threshold to 0.5 to binarise the CAMs, evaluating how pseudo-label construction influences segmentation performance. These ablations reinforce the value of both architectural decisions and preprocessing strategies such as augmentation and adaptive thresholding.

3.3 OEQ: Iterative Self-Training with Superpixel Refinement

To address the OEQ, DeepLabV3 is retrained using its predictions as pseudo-labels, following an initial phase of training on Grad-CAM++ activation maps generated by a ResNet18 classifier. These pseudo-labels are refined using SLIC superpixel segmentation, where each superpixel region is assigned the majority class label from the model's prediction, optionally followed by smoothing. This approach evaluates whether iterative self-learning with superpixel-refined labels can enhance segmentation performance compared to models trained with static Grad-CAM++ masks. The base model for this experiment was the DeepLabV3 trained on Grad-CAM++ pseudo-masks during the MRP phase. This model served as the starting point for the self-learning refinement loop.

4 Results

Table 1. Test set mIoU for different models and training configurations.

Model / Configuration	Epoch 3	Epoch 5	Epoch 10	Epoch 20
Open-ended (Self-learning)	–	–	–	0.7238
Ablation Binary Threshold (DeepLabV3)	–	0.6474	0.6449	–
Ablation No transforms (DeepLabV3)	–	0.6825	0.6865	–
Baseline (DeepLab v3+)	–	0.6917	0.7055	0.6825
U-Net with CAMs	–	0.5708	0.6273	0.6574
PPC Model	0.5433	0.5418	0.5437	0.5437
Fully-supervised U-Net	–	0.7137	0.7792	0.8364

4.1 Fully Supervised vs Weakly Supervised Comparison

The fully supervised U-Net model achieved the best overall performance, with an mIoU of 0.8364 on the test set after 20 epochs, serving as the upper bound for performance. The results for the weakly supervised models are as follows, with DeepLabV+ (10 epochs, thresholds 0.6 and 0.4) achieving the highest mIoU of 0.7055. U-Net, trained on Grad-CAM++ masks, achieved an mIoU of 0.6574, and then its performance decreased. PPC achieved an mIoU of around 0.5437 despite training for up to 20 epochs. This comparison highlights that, although weakly supervised methods do not achieve the high performance of supervised pixel-level training, methods like DeepLabV3 can narrow the gap.

4.2 Ablation Experiments

Training with CAMs binarised at 0.5 achieved mIoUs of 0.6449 (10 epochs) and 0.6474 (5 epochs), underperforming compared to the 0.6/0.4 threshold, suggesting that careful pseudo-label construction is crucial. Removing transformations slightly reduced generalisation, as seen in the improvement from 0.6865 to 0.7055 with augmentation. These results underscore the significance of preprocessing choices, such as augmentation and thresholding, in weakly supervised segmentation.

4.3 OEQ: Self learning + Superpixel Refinement

The self-learning approach further improved DeepLabV3 results: Baseline (no self-learning), 5 epochs of DeepLabV3 with CAMS achieved 0.6917. Self-learning with super pixels and a compactness of 0.01 mIoU improved to 0.7203, and self-learning with super pixels and a compactness of 0.001 mIoU improved to 0.7238. Iterative self-learning with superpixels matched or exceeded performance from longer training (e.g., Baseline DeepLabV3, 10 epochs, mIoU: 0.7055). Thus, self-learning not only improves accuracy but also serves as a robust alternative to pseudo-supervision.

5 Discussion

Grad-CAM++ with DeepLabV3 was the strongest approach, achieving 0.7055 mIoU, further improved to 0.7238 mIoU via self-learning with superpixels. While Grad-CAM++ offers class-discriminative localisation, its coarse resolution can limit precision. Our self-learning pipeline refines these masks iteratively, while PPC tackles the issue by learning contrastive pixel-prototype representations. PPC underperformed by 0.54 mIoU; however, architectural choices, augmentations, and training epochs had a significant impact on the results. Our results show that self-learning improves CAM-based segmentation; however, it remains unclear how the quality of pseudo-labels can be assessed and improved during training without access to ground truth. Another open question is whether the gains from weak supervision on Oxford-IIIT Pet can generalise to more complex datasets or unseen classes. Limitations include testing only on the Oxford-IIIT Pet, limited self-learning, and minimal hyperparameter tuning. Future work could involve additional datasets. CAM refinement and self-training remain promising for WSSS for weakly supervised segmentation.

6 Conclusion

This study explored weakly supervised semantic segmentation using class activation maps, contrastive learning, and iterative self learning. A fully supervised U-Net served as the performance benchmark, while three weakly supervised models, PPC, CAM-based U-Net, and DeepLabV3, were implemented and evaluated. Grad-CAM++, combined with DeepLabV3, consistently outperformed other weakly supervised methods. The best performance (0.7238 mIoU) was achieved through self learning with SLIC superpixel refinement, closely approaching the fully supervised U-Net (0.8364 mIoU). In contrast, PPC based learning showed limited effectiveness despite extended training. Ablation studies highlighted the importance of pseudo-label thresholding, data augmentation, and model architecture. These choices had a measurable impact on the quality of segmentation. Overall, the results demonstrate that structured pseudo labelling with CAMs, combined with architectural design and iterative refinement, can significantly enhance segmentation under weak supervision. This approach provides a robust alternative in settings where pixel-level annotations are unavailable or prohibitively expensive.

References

1. Zhou, Z.H.: A brief introduction to weakly supervised learning. *National Science Review* **5**(1), 44–53 (2018)
2. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *MICCAI 2015, LNCS*, vol. 9351, pp. 234–241. Springer, Cham (2015)

3. Neven, R., Neven, D., De Brabandere, B., Proesmans, M., Goedemé, T.: Weakly-Supervised Semantic Segmentation by Learning Label Uncertainty. arXiv preprint arXiv:2110.05926 (2021)
4. Du, Y., Fu, Z., Liu, Q., Wang, Y.: Weakly Supervised Semantic Segmentation by Pixel-to-Prototype Contrast. arXiv preprint arXiv:2110.07110 (2021)
5. David, L., Pedrini, H., Dias, Z.: P-NOC: Adversarial Training of CAM Generating Networks for Robust Weakly Supervised Semantic Segmentation Priors. arXiv preprint arXiv:2305.12522 (2023)
6. Babenko, B.: Multiple Instance Learning: Algorithms and Applications. University of California, San Diego (2008)
7. Zhu, Y., Zhang, Z., Wu, C., Zhang, Z., He, T., Zhang, H., Manmatha, R., Li, M., Smola, A.: Improving Semantic Segmentation via Efficient Self-Training. arXiv preprint arXiv:2004.08514 (2020)
8. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**(11), 2274–2282 (2012)
9. Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. In: *IEEE Transactions on Image Processing*, vol. 30, pp. 6226–6236 (2021)
10. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv preprint arXiv:1706.05587 (2017)
11. Zou, Y., Zhang, Z., Zhang, H., Li, C.L., Bian, X., Huang, J.B., Pfister, T.: PseudoSeg: Designing Pseudo Labels for Semantic Segmentation. In: *Proceedings of the International Conference on Learning Representations (ICLR)* (2021)
12. Zhu, Y., Zhang, Z., Wu, C., Zhang, Z., He, T., Zhang, H., Manmatha, R., Li, M., Smola, A.: Improving Semantic Segmentation via Efficient Self-Training. arXiv preprint arXiv:2010.09713 (2020)
13. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation. arXiv preprint arXiv:2004.04581 (2020)