

---

# Weakly Supervised Segmentation using Activation Methods

---

Aahn Deshpande

Anish Hota

Harish Karthik Kumaran Pillai

Pulkit Garg

**Mihir Goyenka (Leader)**

Arizona State University

mgoyenka@asu.edu

## Abstract

The creation of masks constitutes a significant manual endeavor, demanding substantial labor resources. In this study, we address this challenge by proposing an automated approach to ameliorate the mask creation process. Our motivation stems from the recognition of the labor-intensive nature of traditional methods. Through our research, we aim to introduce an innovative solution that streamlines mask creation, reducing the burden on human resources while enhancing efficiency. By leveraging image level labels and XAI, we seek to revolutionize the mask-making process, contributing to increased productivity and effectiveness in various applications. Our research question formulated into "Can we use weakly supervised segmentation in coordination with image level labels and activation maps to create masks efficiently and quickly, thereby decreasing the current levels of manual labor and computational requirement that is required?"

## 1 Introduction

In our pursuit to address the labor-intensive nature of manual mask creation, we embarked on a comprehensive exploration, employing a fusion of cutting-edge technologies. Our approach involved the integration of a custom-trained Res-Net based architecture with CCAM, aiming to assess the potential synergies between these frameworks in enhancing mask creation processes. Motivated by the imperative to streamline this labor-intensive task, we sought to investigate the efficacy of this combined approach. Furthermore, drawing from a spectrum of techniques documented in the literature review we engaged in, we conducted a systematic examination, exploring various combinations to ascertain their validity and potential for improvement. Through this concerted effort, we endeavored to pave the way for a more efficient and effective method for mask creation, with implications for diverse applications.

## 2 Literature Review

Class Activation Mapping (CAM), introduced by Zhou et al.[9], is a pioneering technique designed to identify discriminative regions utilized by a restricted class of CNNs, specifically those without fully-connected layers. CAM enhances model transparency by generating visual explanations without requiring architectural changes or re-training, making it a valuable tool for interpreting deep learning models.

Grad-CAM[1] builds upon the principles of CAM to provide class-discriminative localization for any CNN-based network. It achieves this without altering the network's architecture, thus preserving its efficiency and complexity. By leveraging gradients from the final convolutional layer, Grad-

CAM generates heatmaps that highlight the regions influencing the network’s predictions, thereby enhancing interpretability.

HI-RES CAM[7] addresses the limitation of reduced spatial resolution in CAM heatmaps caused by downsampling in the final layers of CNNs. This technique employs strategies such as upsampling the CAM heatmaps or incorporating skip connections to preserve spatial information from earlier layers. By enhancing spatial details, HI-RES CAM provides more accurate and detailed visualizations of the network’s decision-making process. Full-GRAD [6] extends the Grad-CAM algorithm by incorporating feature maps from multiple intermediate convolutional layers. By computing gradients from these layers and employing a fusion mechanism to combine information, Full-GRAD CAM captures both local and global context cues, thereby enriching the interpretability of deep learning models. This comprehensive approach enables a more nuanced understanding of the network’s behavior and decision boundaries.

### 3 Methodology

#### 3.1 Dataset

The experiments were conducted using two distinct datasets selected based on their varying levels of complexity. The first dataset utilized was the Oxford-IIT pet [5] dataset, comprising a total of 7393 images depicting cats and dogs. The second dataset was a composite of two polyp datasets. The training data originated from the TrainingSet\_NewGT dataset comprising 2979 images depicting both polyps and non-polyps. The testing dataset comprised 1000 images sourced from the KVASIR-SEG [3] dataset.

#### 3.2 Architecture

The architecture of the project is illustrated below. All the code for the project can be found in this GitHub repository [2].

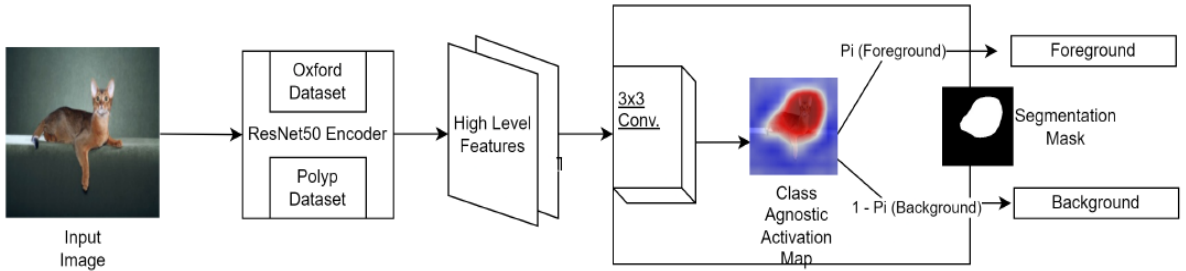


Figure 1: Architecture

##### 1. Encoder Utilization:

In our approach, we leverage ResNet50, a widely acclaimed architecture in computer vision tasks, for feature extraction. By employing ResNet50, we aim to capitalize on its capability to capture intricate visual patterns and representations. Furthermore, to adapt the model to the distinct characteristics of both the Oxford and Polyp datasets, we adopt a dynamic training strategy. This strategy involves alternating training epochs between the two datasets, allowing ResNet50 to adapt iteratively to the specific requirements of each task. Specifically, we conduct 10 epochs of training with image labels to facilitate learning from ground truth annotations, followed by an additional 10 epochs where ResNet50’s weights are frozen, enabling it to serve as a feature extractor. This alternating training regimen enables our model to refine its representations continuously and effectively accommodate the diverse demands of the datasets.

##### 2. Disentangler Architecture:

In the disentangler, we employ a 3x3 convolution layer to commence the process of capturing

local patterns essential for distinguishing foreground from background in the images. Subsequently, we generate class-agnostic activation maps, providing a foundational framework for subsequent processing steps. To delineate foreground and background regions effectively, we adopt a strategy where if a particular activation map  $P_i$  predominantly highlights foreground features, the background regions are derived by computing  $1 - P_i$ . Further refinement of foreground-background representations is achieved through the utilization of flattening and matrix transpose operations on both activation maps and high-level features, ensuring a comprehensive and discriminative representation of the image content. This sequential approach ensures a systematic extraction and refinement of crucial visual cues, facilitating accurate segmentation and analysis of foreground and background elements in the images.

3. **Modification of CCAM Architecture:** Adapt CCAM [8] to integrate class labels for potentially improved performance.
4. **Comprehensive Methodology:** Integration of ResNet50 and novel disentangler for robust feature representation and segmentation in complex image datasets.
5. **Other activation maps for benchmarking:** In addition to employing the original CCAM and the modified architecture outlined above, our study encompassed the utilization of alternative activation methods for comparative analysis. Specifically, we explored the efficacy of FullGrad, HiResCAM, and GradCAM in our investigations. GradCAM operates by extracting gradients from the final fully convolutional layer and subsequently generating a heatmap. Conversely, HiResCAM diverges from GradCAM by refraining from averaging the gradients, opting instead to multiply the entire gradient matrix with the input image. This divergence holds potential for yielding superior results, particularly in scenarios such as the polyp dataset, where capturing intricate pixel features is imperative, as opposed to the oxford dataset. On the other hand, FullGrad generates activation maps by amalgamating activation maps derived from gradients across all intermediate layers. This comprehensive approach allows for a nuanced understanding of the underlying data, potentially enhancing the depth of insights gleaned from the analysis.

## 4 Results

We use the Dice coefficient and IoU metrics in addition to the other metrics primarily because both these metrics are commonly used in image segmentation tasks. Also, they are robust to class imbalance because they evaluate the overlap between the predicted and ground truth segmentation masks for each class separately.

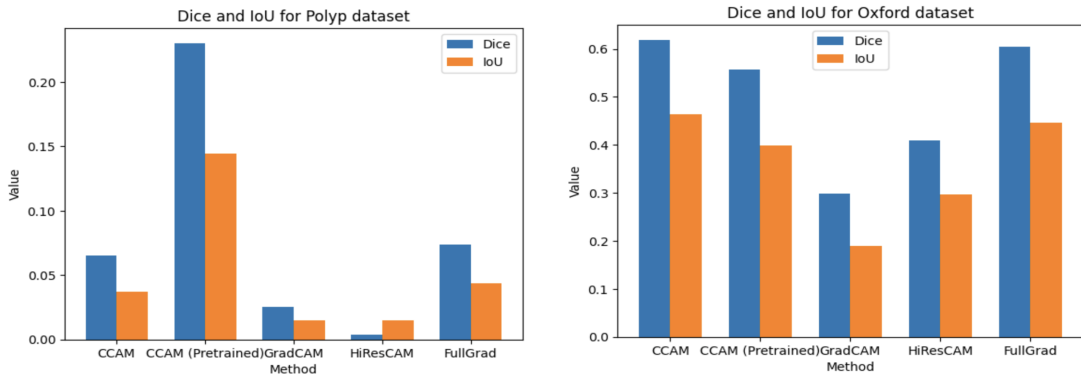


Table 1: Oxford and Poly Metrics

## 5 Discussion

1. **Enhanced Segmentation Mask Generation for Polyps through Combined Image-Level Supervision and Foreground/Background Features:**  
Our initial attempts at feature extraction encountered hurdles specific to the polyp dataset.

Unlike the Oxford dataset, where foreground and background features were readily discernible, the absence of clear boundaries in polyp images posed significant challenges. Traditional methods struggled to accurately delineate between polyp regions and their surroundings. In response, we explored leveraging a ResNet model with image-level supervision as an encoder prior to feature extraction. This approach proved pivotal in overcoming the inherent complexities of the polyp dataset. By incorporating high-level semantic information through image-level supervision, the ResNet model facilitated more robust feature extraction. Consequently, we observed superior performance compared to conventional methods, underscoring the importance of adapting feature extraction strategies to the unique characteristics of the dataset.

## 2. Insufficient Improvement with Activation Maps Utilizing Intermediate Layers:

Initially, our strategy entailed utilizing gradients from intermediate layers to enhance the activation map, surpassing Grad-CAM’s reliance solely on the final layer of the CNN. Our aim was to merge concept activation vectors with Grad-CAM activation maps to optimize our approach. However, our attempts to integrate the Testing with Concept Activation Vectors (TCAV) [4] framework with activation maps encountered technical obstacles, particularly regarding the compatibility between PyTorch code and TCAV’s original TensorFlow implementation. Despite our concerted efforts to reconcile these disparate components, compatibility issues persisted, hindering seamless integration. Consequently, we explored alternative methodologies to achieve comparable functionality. One such approach involved adopting the FullGrad activation method, which computes gradients across all layers of the network. This alternative method provided a practical solution to circumvent the limitations encountered with TCAV integration. By harnessing FullGrad, we were able to attain similar functionality and successfully fulfill our intended objectives. However, FullGrad didn’t perform better than our CCAM architecture in the polyp dataset.


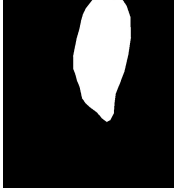
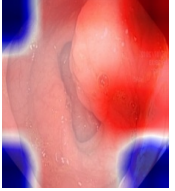

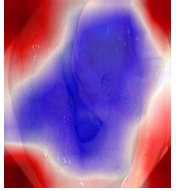


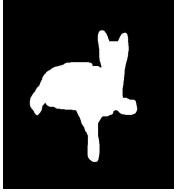
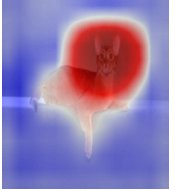

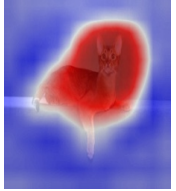

Polyp Dataset					
Original & Ground Truth:		CCAM pretrained:		CCAM Original	
					
Oxford Dataset					
Original & Ground Truth:		CCAM pretrained:		CCAM Original	
					

Table 2: Images after training

## 6 Conclusion & Future Work

From the findings, it appears that weakly supervised segmentation presents a promising alternative to manual mask creation through human annotation. Our observations indicate that employing a hybrid approach, combining image-level supervision with a discernment of foreground and background features, yields notably superior results. With the evolution of self-supervised learning techniques and the emergence of foundational models like SAM or MedSAM, significant strides have been made in the realm of mask creation without human intervention. These models exhibit heightened accuracy and have the potential to streamline image labeling processes. Furthermore, beyond the methodologies investigated in this study, there exist additional weakly supervised segmentation techniques that warrant exploration for benchmarking purposes. Additionally, the modified architecture could be subjected to evaluation using larger datasets such as COCO or VOC 2012, enabling comparative analysis against alternative models.

## References

- [1] Hang-Cheng Dong, Yuhao Jiang, Yingyan Huang, Jingxiao Liao, Bingguo Liu, Dong Ye, and Guodong Liu. Rethinking class activation maps for segmentation: Revealing semantic information in shallow layers by reducing noise, 2023.
- [2] Anish Hota and Mihir Goyenka. Weakly Supervised Segmentation using Activation maps. <https://github.com/AnishHota/weakly-supervised-segmentation.git>.
- [3] Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D. Johansen. Kvasir-seg: A segmented polyp dataset, 2019.
- [4] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *arXiv e-prints*, page arXiv:1711.11279, November 2017.
- [5] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [6] Francois Fleuret Suraj Srinivas. Full-gradient representation for neural network visualization, 2019.
- [7] Thanos Tagaris, Maria Sdraka, and Andreas Stafylopatis. High-resolution class activation mapping. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4514–4518, 2019.
- [8] Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, and Linlin Shen. Contrastive learning of Class-agnostic Activation Map for Weakly Supervised Object Localization and Semantic Segmentation. *arXiv e-prints*, page arXiv:2203.13505, March 2022.
- [9] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015.