Better Image Segmentation with Classification: Guiding Zero-Shot Models Using Class Activation Maps

Hanna Borgli[1,3(✉)], Håkon Kvale Stensland[1,3], and Pål Halvorsen[2]

[1] Universitetet i Oslo, Oslo, Norway
[2] SimulaMet, Oslo, Norway
paalh@simula.no
[3] Simula Research Laboratory, Oslo, Norway
{hanna,haakonks}@simula.no

Abstract. In this demo paper, we present a method for image segmentation of images with no segmentation masks available and a web application built on top of it. The method does not require any segmentation images for training; instead, it uses a classifier trained on classification images to power a zero-shot method to get the segmentation masks. We focus on gastrointestinal images and include datasets with both ground truth masks available and not available. In this demo, the user is shown the steps for generating the mask and can observe where the method excels and is challenged. We also allow the users to experiment with parameters for the generation but have provided default parameters based on experiments. With this system, users can get masks for any object for which classification data is available. They can use it to test the viability of creating hand-crafted masks for training, automatically annotate datasets, and deploy it as-is. A video demonstration can be found at https://youtu.be/YjX19bBXf6Y.

Keywords: Image Segmentation · Class Activation Map · Segment Anything · Annotation Tools · Zero-Shot Learning

1 Introduction

In recent years, foundational models have been released to create image segmentation masks for general objects [9,12]. They are trained on large datasets and can perform zero-shot for most objects. In medicine, these general models can be helpful in cases where we do not have data available, as medical data can be expensive to collect and have many restrictions. In this paper, we will focus on gastrointestinal images for our data.

Several works have explored utilizing zero-shot models such as Segment Anything (SAM) [9] for polyp segmentation [2,10,11]. The capabilities of the zero-shot models are either prompt-based mask generation using prompts such as text, bounding boxes, masks, or points indicating foreground and background,

I. Ide et al. (Eds.): MMM 2025, LNCS 15524, pp. 105–111, 2025.
https://doi.org/10.1007/978-981-96-2074-6_10

or automatic-based, where the model tries to find all distinct objects in an image. Therefore, works are usually based on fine-tuning the zero-shot model, manipulating the prompts, or filtering the automatically generated masks.

In this demo, we will present a web application for generating segmentation masks for gastrointestinal images. Our method uses class activation maps (CAMs) to produce bounding boxes that serve as prompts for the zero-shot models. The resulting mask is compared with a set of masks generated using the automatic generation functionality of the zero-shot model. The mask with the highest intersection over union (IoU) score relative to the CAM-generated mask is selected if it has an acceptable IoU score. If none is found, we use the CAM-generated mask.

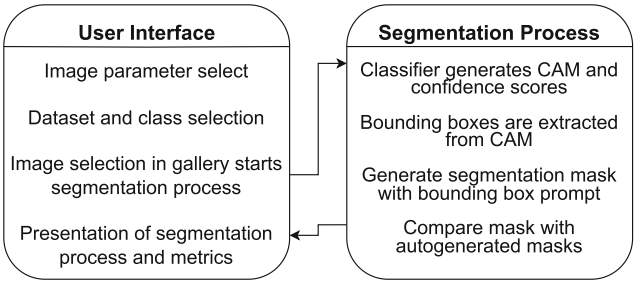


Fig. 1. The figure presents an overview of the system, comprising a Gradio-based web interface and a segmentation process. The interface enables users to set segmentation parameters like CAM method and bounding box threshold and choose a dataset, class, and image to initiate segmentation. The process involves generating the CAM, extracting the bounding box, and producing a segmentation mask with a zero-shot model. Results are evaluated by comparing the mask against automatically generated masks using IoU. If IoU exceeds 0.1, the overlap is displayed; otherwise, the generated mask is shown.

2 System Architecture

The system uses a web interface front-end and a Python server running the method. The Python library Gradio [1] is used for the server and web interface. The method uses the ML library Pytorch and Pytorch CAM [5] to generate CAMs. Further, we use SAM and SAM 2 depending on the user selection for the zero-shot method. Figure 1 shows how the system is designed and each step of the segmentation mask process. When we evaluate the masks, we use IoU, a standard metric used for mask segmentation evaluation and explained in related work such as the dataset papers [6,8]

2.1 Datasets

In this demo, user access is confined to pre-selected gastrointestinal image datasets. These include polyp images from Kvasir-SEG [8], instrument images from Kvasir-Instruments [6], and training and testing splits from GastroVision [7], which were employed in classifier training. This selection allows users to compare training data performance versus unseen data. The focus on gastrointestinal images aligns with our testing and classifier training processes.

- **Kvasir Segmented Polyps (Kvasir-SEG)** [3,8] Focused on polyp detection in gastrointestinal endoscopic exams, Kvasir-SEG contains 1,000 images with ground truth masks. Originating from the Kvasir dataset, it's a benchmark for evaluating polyp segmentation models, motivating its inclusion in this paper.
- **Kvasir Instruments (Kvasir-instruments)** [6] This dataset focuses on instruments found in gastrointestinal tract examinations, featuring 590 masks, and complements the Kvasir-SEG dataset.
- **GastroVision** [7] With 8,000 images from 27 gastrointestinal classes, sourced from two hospitals in Norway and Sweden and annotated by clinicians, GastroVision supports image classification. The dataset includes splits for training, testing, and validation, utilized for training the classifier and for demonstration datasets in this paper.

2.2 Zero-Shot Models

The Segment Anything Model [9], released by Meta, offers us a way of creating segmentation masks based on prompts such as bounding boxes, masks, and selected pixels. Much of the utility behind SAM is its ability to take different custom prompts to create segmentation masks. SAM has two modes: (1) The predictor mode, where the model creates a mask based on a prompt, and (2) the automatic mask generator mode, where the model creates masks for all objects in the image based on a point grid. The automatic mask generator also has different parameters affecting the generation of masks, but for this demo, we use the default parameters. There is also a newer version of SAM called Segment Anything 2 (SAM 2) [12]. SAM 2 improves SAM's general mask generation capabilities and allows for video mask generation. For our demo, we only use the image capabilities, and the user can select which version of SAM they want to use.

2.3 Mask Generation

Our system for mask generation requires a classification model and a zero-shot model. We trained the classification model in the same way described in the dataset paper of GastroVision [7] using DenseNet-121. The zero-shot model was either SAM or SAM 2 based on the user's selection. The user can select between the CAM methods used and enable smoothing for the CAM. Eigen smoothing

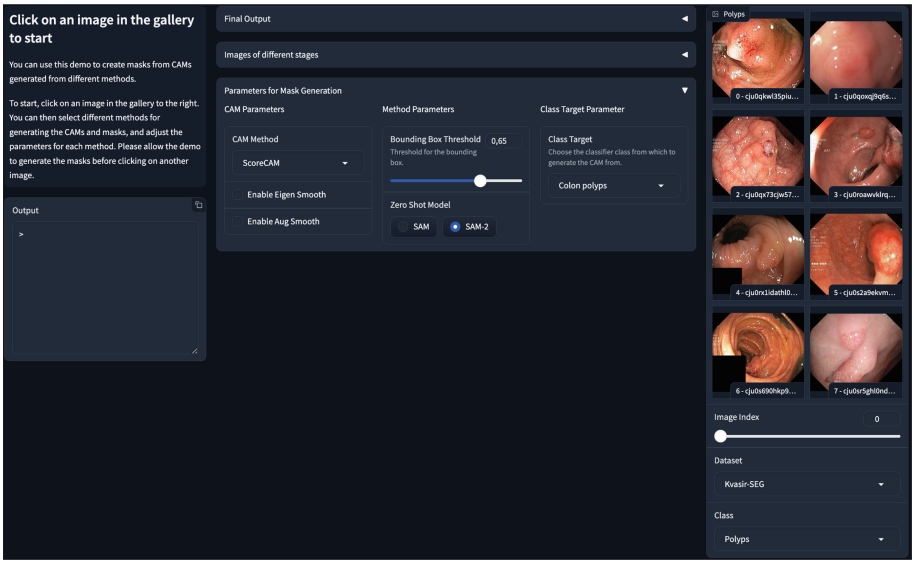


Fig. 2. The landing page of the demo. The page is split into three columns. The first from the left (1), shows text explaining how to start the demo for the user and under there is an output box which will show output from the mask generation. The middle column (2), has three sections, but only the first is open when the user enters. In this section, the user can select parameters for running the mask generation. The other sections present the generated mask and images showing steps of the mask generation. The third column (3), shows a gallery view of the currently selected dataset and class and the page index of those images. When the user clicks one of the images, the generation starts.

filters out noise in the CAM, and augmentation smoothing centers the CAM more around the object. Augmentation smoothing especially increases the time it takes to generate a CAM. The CAM methods available are the same as those in the Pytorch CAM [5] Python package. Further, the user can select which class they want the CAM to target. This can be useful if there are more classes in one image and the user wants to target one of the other classes.

When an image is selected, a CAM is generated based on the user's selected parameters. The following heat map is then used to extract a bounding box. This is based on the bounding box threshold, the last parameter the user can select. A threshold of 0.65 means that a bounding box is extracted, including all values of 0.65 and over. So, a bigger threshold will equal a smaller, more targeted bounding box. Users should experiment with different sizes as some classes might benefit from more targeted bounding boxes. The demo's default parameters, ScoreCAM with no smoothing enabled and 0.65 bounding box threshold, were found through experiments comparing different parameters for the Kvasir-SEG dataset.

The bounding box is used as a prompt for the predict functionality of SAM. The resulting mask is then compared to the list of images generated using SAM's

automatic mask generator functionality. The list of images is first pruned for masks that are too big, which is typical when the whole viewport is made into a mask or encompasses image artifacts such as the borders. We use the IoU score to determine which mask in the list best overlaps with our CAM-generated mask. The best overlap is chosen if the IoU score is higher than 0.1 to ensure the overlap is good enough. If we can not find any overlap or the overlap does not produce an IoU over 0.1, we present the CAM-generated mask as the final result.

3 Demo

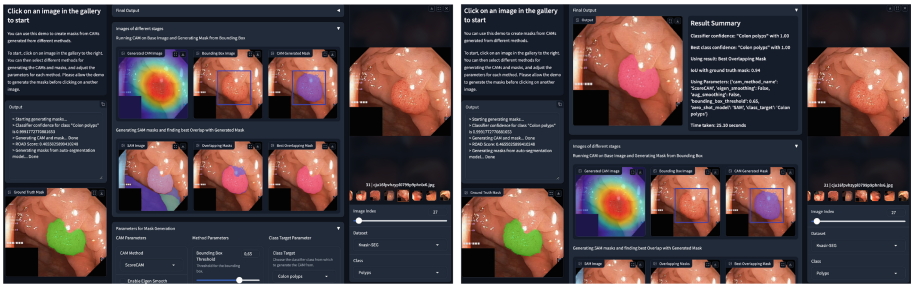
In this demo paper, we present a web user interface utilizing our method of extracting bounding boxes from class activation maps and using them as inputs for zero-shot segmentation models. The landing page for the demo can be seen in Fig. 2, and is designed to use three columns. The left column is the information column. Here, we have a little section of text quickly explaining how the demo works and how to get started. Under this, we have an output box that starts empty. The box will fill with text when the mask generation is starting, informing the user of where in the generation process the system is and some metrics such as the confidence score of the classifier prediction for the selected class and the remove and the debias [4] metric of the CAM.

The central column is the configuration and results column. There are three sections: (1) The final output section, containing the final output mask overlaid on the original image and some metrics such as the confidence score. (2) The images from different stages section, containing six images showing the different stages of the mask generation. (3) The generation parameters section contains adjustable parameters, which affect the result of the segmentation mask generation step.

The parameters selection section has the following parameters to select from:

- CAM parameters: Selection contains a drop-down menu of the different CAM methods, where only one can be selected at a time and two selection boxes that can be toggled, turning on or off Eigen smoothing and augmentation smoothing.
- Mask generation parameters: Selection contains a slider setting the bounding box threshold, which can be a value between 0 and 1, and a radio selection box allowing either SAM or SAM 2 to be used as the zero-shot model to be used in the generation steps.
- Class target: Sets the class target we use to generate the CAM. For most images, this should stay default as it will automatically be set to the class the image is classified to, but other classes are present for some images. We therefore include it as an option.

The final column is the gallery column. Here, the user can use a slider to view the images in the selected class and dataset. They may also change the dataset and class among the datasets and classes included in the demo. When



(a) (b)

Fig. 3. The combined figure shows the stage of image generation to the left and the final stage of presenting the resulting mask on the right. When the image is being generated, there are orange outlines on web components affected by the generation, and the images showing the steps of the generation are automatically shown. When the mask generation is done, the result page is extended automatically, and we are presented with the final mask and a summary of the results. (Color figure online)

a user clicks an image in the gallery, the image is shown, and the segmentation mask generation starts.

When the generation starts, the parameter section closes, and the images of the different stages section open. Additionally, we have two different behaviors depending on whether the dataset can access ground truth masks. If there is no ground truth mask, the output box fills with text informing the user of each step. Further, the images in the different stages section appear as they are made. In the end, the final output section is opened, and the final segmentation mask is presented along with the confidence score for the class target, the best confidence score, and the class target it was achieved with, whether we are using the bounding box generated mask or an automatically generated mask with good overlap, which parameters were used for the generation and the total time taken. If there is a ground truth mask, we present that mask at the bottom of the first column. This is shown in both figures in Fig. 3. Furthermore, we present the IoU score between the final resulting mask and the ground truth.

4 Conclusion

This demo paper presents our web-based system for generating segmentation masks using CAMs and the zero-shot model SAM. By training a classifier and extracting bounding boxes from generated CAMs, we can guide the zero-shot models into generating better results, which is especially useful for objects where we do not have segmentation masks available for training. We have created a web interface where users can select images between different datasets and classes and where the user can adjust different parameters that can affect the results. The demonstration shows a system that can annotate and understand masks for

objects that do not have masks for training data. It can also be used as-is in other systems to increase the performance of zero-shot foundational models.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abid, A., Abdalla, A., Abid, A., Khan, D., Alfozan, A., Zou, J.: Gradio: hassle-free sharing and testing of ML models in the wild (2019). arXiv preprint arXiv:1906.02569 https://arxiv.org/abs/1906.02569
2. Biswas, R.: Polyp-SAM++: can a text-guided SAM perform better for polyp segmentation? arXiv preprint arXiv:2308.06623 (2023). https://arxiv.org/abs/2308.06623
3. Borgli, H., et al.: HyperKvasir: a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Sci. Data **7**(1), 283 (2020)
4. Chu, S., Kim, D., Han, B.: Learning debiased and disentangled representations for semantic segmentation. Adv. Neural. Inf. Process. Syst. **34**, 8355–8366 (2021)
5. Gildenblat, J., contributors: Pytorch library for cam methods (2021). https://github.com/jacobgil/pytorch-grad-cam
6. Jha, D., et al.: Kvasir-Instrument: diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In: Proceedings of the 27th International Conference on Multimedia Modeling (MMM), pp. 218–229. Springer, Cham (2021)
7. Jha, D., et al.: GastroVision: a multi-class endoscopy image dataset for computer-aided gastrointestinal disease detection. In: ICML Workshop on Machine Learning for Multimodal Healthcare Data (ML4MHD 2023) (2023)
8. Jha, D., et al.: Kvasir-SEG: a segmented polyp dataset. In: Proceedings of the 26th International Conference on Multimedia Modeling (MMM), pp. 451–462. Springer (2020)
9. Kirillov, A., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)
10. Li, Y., Hu, M., Yang, X.: Polyp-SAM: transfer SAM for polyp segmentation. In: Medical Imaging 2024: Computer-Aided Diagnosis. vol. 12927, pp. 759–765. SPIE (2024)
11. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nat. Commun. **15**, 1–9 (2023)
12. Ravi, N., et al.: SAM 2: segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024). https://arxiv.org/abs/2408.00714