Weakly supervised automated extraction of optimal visual prompts for foundational models in medical image segmentation

Vivek Dhamale & Rushikesh Pawar
Department of Computational and Data Sciences
Indian Institue of Science Bangalore
Karnataka, India 560 012
{viveksd,rushikeshp}@iisc.ac.in

Abstract

The lack of publicly available quality annotations in medical imaging has been the bottleneck for training large-scale deep learning models for medical image segmentation, due to the time-consuming nature of the task. Vision foundation models, often trained on large-scale visual images of various modalities, could serve as the basis for building medical segmentation applications. Promptable vision foundational model SAM is pretrained on natural images to return valid segmentation mask for a relevant prompt, (e.g., foreground/background points, a rough box or mask, freeform text, or, even text, indicating the regions to segment in an image). However, SAMs typically involve providing manual prompts. In this paper we propose a method to automatically extract optimal prompts for SAM is a weakly supervised manner. We also evaluate the performance of different prompts (point,bounding box and their combination) on the task of medical image segmentation. We compare the quality of automatically extracted prompts with manually provided prompts on the basis of how well they perform on medical image segmentation task. The code is available at `https://github.com/rushikeshpawar22581/DLCV-Project2024-Automated_prompt_extraction_for_SAM`

1 Introduction

Medical image segmentation is a crucial task in medical image analysis, as it provides a detailed understanding of the anatomical structures present in the image. Deep learning models have shown promising results in medical image segmentation tasks, but they require a large amount of annotated data for training. However, the lack of publicly available quality annotations in medical imaging has been the bottleneck for training large-scale deep learning models for medical image segmentation, due to costly and the time-consuming nature of the task. One way to overcome this limitation is to use vision foundation models, often trained on large-scale visual images of various modalities, as the basis for building medical segmentation applications. Promptable vision foundational model SAM [Kirillov et al., 2023] (Segment Anything Model, developed by Meta) is pretrained on natural images to return valid segmentation mask for a relevant prompt, (e.g., foreground/background points, a rough box or mask, freeform text, or, even text, indicating the regions to segment in an image). This prompt is usually provided manually. Which too requires expert's time and effort. We propose a method to automatically extract optimal prompts for SAM in a weakly supervised manner. We use classification labels to train a classifier and use CAM to extract the optimal prompt for the segmentation task. We also evaluate the performance of different prompts (point, bounding box and their combination) on the task of medical image segmentation on BUSI dataset [Al-Dhabyani et al., 2020]. We compare

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

the quality of automatically extracted prompts with prompts generated from ground truth on tumor segmentation task of BUSI dataset.

2 Related work

In the study [Hu et al., 2023], authors used SAM for segmentation of breast tumors in ultrasound images. They highlighted significance of prompt interaction in improving the model's segmentation performance, with substantial improvements in performance metrics when prompts were incorporated.

In another research work [Ma et al., 2024], authors fine-tuned SAM on a dataset with more than one million medical image-mask pairs. They call it MedSAM. MedSAM also requires manual prompts for segmentation task.

Wu et al. [2023] proposed a self-prompting method in medical vision applications. They harnessed the embedding space of SAM to prompt itself through a linear pixel-wise classifier. Their approach requires access to ground truth masks while training the self-prompting module. Which is not always available in medical imaging datasets.

A recent study [Na et al., 2024] propose an auto-prompt generator. Their approach is to train an auxiliary neural network (They used U-net in the paper) to predict the binary mask of the objects of interest. They used the predicted mask as a prompt for SAM. This is also a supervised approach and requires ground truth masks for training the auxiliary network.

Study by [Cui et al., 2024] proposed two stages pipeline: SAM empowered annotation and SAM finetuning. In the first stage,They provide the bounding boxes of nuclei as input to the pretrained SAM model to generate high-quality approximate nuclei masks. the second stage, the generated approximate masks are employed to finetune the SAM model. Even though this counts as weakly supervised approach, it still requires manual bounding boxes as input.

As mentioned above, most of the existing methods for automated prompt generation require ground truth masks or bounding boxes as input. We propose a method in which we use classification labels to train a classifier and use CAM to extract the optimal prompt for the segmentation task on BUSI dataset. We also evaluate the performance of different prompts (point, bounding box and their combination) on the task of medical image segmentation on BUSI dataset. We compare the quality of automatically extracted prompts with prompts generated from ground truth on tumor segmentation task of BUSI dataset as done in [Hu et al., 2023].

3 Methodology:

The methodology that we have followed can be divided into two parts; Extracting visual prompts from class activation maps and using SAM model to predict the segmentation masks.

3.1 Generating Class activation Maps:

For extracting accurate visual prompts we have used 2 different frameworks: binary classification of normal and abnormal (Benign and malignant) classes and 3 class classification among normal, benign and malignant class. For classification we have experimented with models from ResNet family (ResNet 18, ResNet 34, ResNet 50). ResNet 18 and ResNet 50 were used for prompt extraction after 3 class classification while ResNet 34 was used in binary classification setup. We have used weighted cross entropy loss. Weights for different classes were tuned manually to get better performance in F1 score as well as recall for critical classes. In all of models we have attached extra classification head (a non-linearity (ReLU), a dropout layer with and single FC layer).Models were trained with heavy regularization (0.7 dropout with weight decay too).We have initialized all models with Image-Net weights and then all layers were fine-tuned on BUSI[Al-Dhabyani et al., 2020] dataset. Total dataset was split into 80% train and 20% validation set. After training both type of classification models, CAM extraction was done by using GradCAM [Selvaraju et al., 2017] and GradCAM++ for benign and malignant class images.[Chattopadhay et al., 2018]. We have used TorchCAM library for utilizing CAMs in this projectFernandez [2020]

3.2 Extraction of Visual prompts:

After getting CAMs from classification model, we threshold CAM to generate pseudo mask for segmentation. We extracted 3 different kind of prompts from pseudo mask Box prompt, Box + Point Prompt and Point prompt. All point prompts refers to foreground point. We have experimented with center point as well as maximum saliency value point within the mask (MaxPoint) as a point prompt.

3.3 Using SAM model to predict the segmentation masks:

SAM model comes with 3 different encoders: ViT-B (91M), ViT-L (308M), and ViT-H (636M). As per BreastSAM, ViT-L performs best, Hence we use ViT-L (308M). Visual Prompts extracted from above method are used to predict the segmentation mask using SAM model.

3.4 Evaluation methodology:

We have evaluated the performance of our method using validation dataset using metrics like Intersection over union (IoU), Dice score and pixel wise accuracy and compared its performance with Breast-SAM methodology where they used ground truth to generate visual prompts.We used ground truth mask with center-point as point prompt and bounding box of mask as box prompt. We didn't used random shift and scale as used in Breast-SAM paper since we wanted to get best performance of SAM for these prompts.

Evaluation Criterion	Formula
Pixel Accuracy	Number of correctly predicted pixels/ Total number of Pixels
IoU (Jaccard)	Intersection/ (Prediction + Ground Truth - Intersection
Dice Score	2 * Intersection / (Prediction + Ground Truth)

4 Results and Discussion

4.1 Three-class classification performance across ResNet models

Table 1 shows the performance of ResNet models on three-class classification task.

Table 1: Three-class classification performace across ResNet models

	Precision			Recall			F1 score		
	ResNet18	ResNet34	ResNet50	ResNet18	ResNet34	ResNet50	ResNet18	ResNet34	ResNet50
Normal	0.83	0.81	0.78	0.83	0.86	0.72	0.83	0.83	0.75
Benign	0.87	0.86	0.84	0.89	0.92	0.9	0.88	0.89	0.87
Malignant	0.79	0.92	0.82	0.75	0.75	0.73	0.77	0.82	0.77

Further table 2 shows the comparison of segmentation performance on held out validation set using GradCAM generated prompts on ResNet18. Percentage of segmentation performance using GT-promt is given in bracket. Bold letters indicate more than 50% performace of that of GT-prompt performace for DICE and IOU and more than 80% for pixel accuracy.

Similary 3 shows the comparison of segmentation performance on held out validation set using GradCAM++ generated prompts on ResNet18. It can be observed that GradCAM performs better than GradCAM++ in generating prompts for medical image segmentation task.

In table 4 we compare the performance of ResNet50 Classifier with GradCAM vs GradCAM++ generated prompts. We can observe that GradCAM performs better than GradCAM++ here as well. (Bold letter indicates better performance and percentage of GT-prompt performance is given in bracket)

Table 5 shows the comparison of segmentation performance on held out validation set using ResNet18 vs ResNet50 using GradCAM. We can observe that ResNet18 performs better than ResNet50 in generating prompts for medical image segmentation task. (Percentage of GT-prompt performance is given in bracket)

Table 2: Comparison of segmentation performance on held out validation set using GradCAM generated prompts on ResNet18.

Type of Tumor	Prompt	Average Dice Score		Average IOU		Average Pixel Accuracy	
		Prompts from GradCAM	Prompts from Ground Truth	Prompts from GradCAM	Prompts from Ground Truth	Prompts from GradCAM	Prompts from Ground Truth
Benign	Box	0.37 (43%)	0.86	0.30 (39%)	0.78	**0.94 (95%)**	0.98
Malignant	Box	0.32 (39%)	0.82	0.23 (33%)	0.71	**0.87 (91%)**	0.95
Benign	Box + 1 Point	0.33(38%)	0.87	0.27 (34%)	0.79	**0.93 (94%)**	0.98
Malignant	Box + 1 Point	0.33(40%)	0.83	0.24 (34%)	0.72	**0.87 (91%)**	0.96
Benign	1 Point	**0.40 (59%)**	0.67	**0.35 (59%)**	0.59	0.67 (75%)	0.90
Malignant	1 Point	**0.35 (59%)**	0.58	**0.27 (59%)**	0.46	0.65 (65%)	0.81

Table 3: Comparison of segmentation performance on held out validation set using ResNet18 and GradCAM++ generated prompts.

Type of Tumor	Prompt	Average Dice Score		Average IOU		Average Pixel Accuracy	
		Prompts from GradCAM ++	Prompts from Ground Truth	Prompts from GradCAM ++	Prompts from Ground Truth	Prompts from GradCAM ++	Prompts from Ground Truth
Benign	Box	0.32 (37%)	0.86	0.25 (33%)	0.78	**0.94 (95%)**	0.98
Malignant	Box	0.17 (20%)	0.82	0.12 (17%)	0.71	**0.87 (91%)**	0.95
Benign	Box + 1 Point	0.25 (28%)	0.88	0.19 (24%)	0.79	**0.93 (94%)**	0.98
Malignant	Box + 1 Point	0.18 (21%)	0.83	0.13 (18%)	0.72	**0.87 (91%)**	0.96
Benign	1 Point	**0.18 (26%)**	0.68	**0.15 (26%)**	0.59	0.67 (75%)	0.90
Malignant	1 Point	**0.25 (43%)**	0.59	**0.18 (40%)**	0.46	0.65 (80%)	0.81

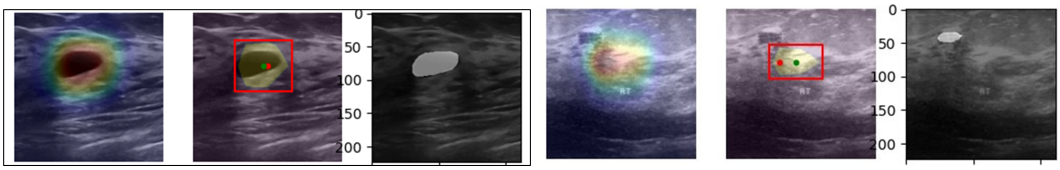
Table 4: Comparison of segmentation performance on held out validation set using ResNet50 Classifier with GradCAM vs GradCAM++.

Type of Tumor	Prompt	Average Dice Score		Average IOU		Average Pixel Accuracy	
		Prompts from GradCAM	Prompts from GradCAM ++	Prompts from GradCAM	Prompts from GradCAM ++	Prompts from GradCAM	Prompts from GradCAM ++
Benign	Box	**0.32 (37%)**	0.26 (30%)	**0.25 (32%)**	0.19 (25%)	**0.93 (94%)**	0.93 (92%)
Malignant	Box	**0.32 (39%)**	0.31 (37%)	**0.23 (32%)**	0.22 (30%)	**0.87 (91%)**	0.87 (86%)
Benign	Box + 1 Point	**0.20 (23%)**	0.17 (20%)	**0.13 (17%)**	0.12 (15%)	**0.91 (93%)**	0.91 (91%)
Malignant	Box + 1 Point	**0.34 (41%)**	0.33 (39%)	**0.24 (34%)**	0.23 (32%)	**0.87 (91%)**	0.87 (86%)
Benign	1 Point	0.12 (18%)	**0.15 (22%)**	0.10 (16%)	**0.12 (21%)**	**0.51 (57%)**	0.51 (46%)
Malignant	1 Point	**0.36 (60%)**	0.35 (60%)	0.26 (57%)	0.26 (56%)	**0.58 (71%)**	0.58 (62%)

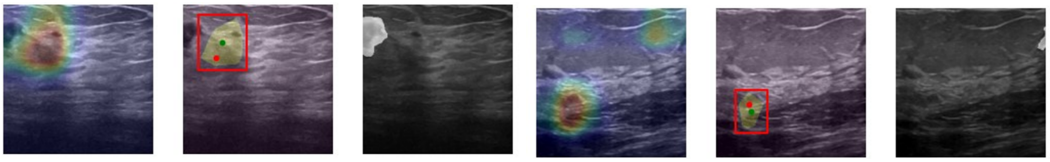
Table 5: Comparison of segmentation performance on held out validation set using ResNet18 vs ResNet50 using GradCAM. (Bold letter indicates performace is greater than or equal to 5%)

Type of Tumor	Prompt	Average Dice Score		Average IOU		Average Pixel Accuracy	
		ResNet18	ResNet50	ResNet18	ResNet50	ResNet18	ResNet50
Benign	Box	**0.37 (43%)**	0.32 (37%)	**0.30 (39%)**	0.25 (32%)	0.94 (95%)	0.93 (94%)
Malignant	Box	0.32 (39%)	0.32 (39%)	0.23 (33%)	0.23 (32%)	0.87 (91%)	0.87 (91%)
Benign	Box + 1 Point	**0.33(38%)**	0.20 (23%)	**0.27 (34%)**	0.13 (17%)	0.93 (94%)	0.91 (93%)
Malignant	Box + 1 Point	0.33(40%)	0.34 (41%)	0.24 (34%)	0.24 (34%)	0.87 (91%)	0.87 (91%)
Benign	1 Point	**0.40 (59%)**	0.12 (18%)	**0.35 (59%)**	0.10 (16%)	**0.67 (75%)**	0.51 (57%)
Malignant	1 Point	0.35 (59%)	0.36 (60%)	0.27 (59%)	0.26 (57%)	0.65 (65%)	**0.58 (71%)**

4.2 Visualizing prompts and CAMs:

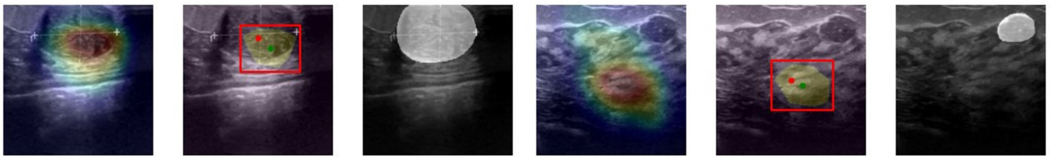


(a) Prediction: 1 GT: 1 (b) Prediction: 2 GT: 1

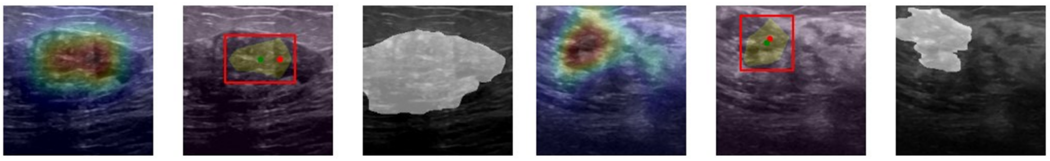


(c) Prediction: 2 GT: 2 (d) Prediction: 1 GT: 2

Figure 1: ResNet 18 GradCAM ++ with in each set of 3 figures. Left figure as CAM, middle as Psuedo mask with prompts and right as ground truth mask, 0 : normal, 1: benign, 2: malignant



(a) Prediction: 1 GT: 1 (b) Prediction: 0 GT: 1



(c) Prediction: 2 GT: 2 (d) Prediction: 1 GT: 2

Figure 2: ResNet 18 GradCAM ++ with in each set of 3 figures. Left figure as CAM, middle as Psuedo mask with prompts and right as ground truth mask, 0 : normal, 1: benign, 2: malignant

4.3 Binary Classifier

Table 6 shows the performance of ResNet34 binary classifier on the task of classifying normal and abnormal images. We can observe that the classifier performs well on the task of binary classification, achieving high precision, recall and F1 score for abnormal class.

Table 6: ResNet34 binary classifier performance

	Precision	Recall	F1 score
Normal	0.84	0.90	0.87
Abnormal	0.98	0.96	0.97

Table 7 shows the comparison of segmentation performance on held out validation set using CenterPoint and MaxPoint for ResNet34 binary classifier with GradCAM. It can be observed that CenterPoint performs better than MaxPoint in generating prompts for medical image segmentation task. (Percentage of GT-prompt performance is given in bracket)

Table 8 shows the comparison of segmentation performance on held out validation set for ResNet18 3-class classifier and ResNet34 2-class classifier using GradCAM. It can be observed that ResNet34

Table 7: Comparison of segmentation performance on held out validation set using CenterPoint and MaxPoint for ResNet34 binary classifier with GradCAM

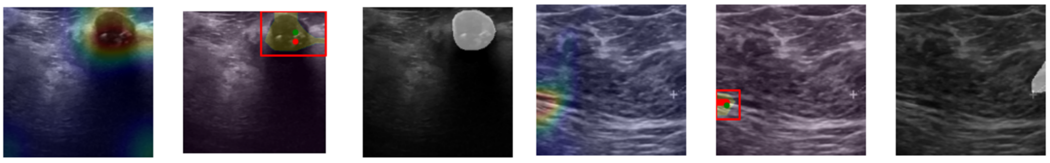
Type of Tumor	Prompt	Avg Dice Score		Avg IOU		Avg Pixel Accuracy	
		CenterPoint	MaxPoint	CenterPoint	MaxPoint	CenterPoint	MaxPoint
Benign	Box	**0.47 (54%)**	0.35 (40%)	**0.39 (50%)**	0.27 (35%)	0.94 (96%)	0.89 (90%)
Malignant	Box	0.17 (20%)	0.18 (21%)	0.12 (17%)	0.13 (18%)	0.83 (87%)	0.77 (80%)
Benign	Box + 1 Point	**0.46 (53%)**	0.34 (39%)	**0.38 (49%)**	0.27 (34%)	0.94 (96%)	0.88 (89%)
Malignant	Box + 1 Point	0.16 (20%)	0.19 (23%)	0.12 (16%)	0.14 (19%)	0.83 (87%)	0.76 (80%)
Benign	1 Point	**0.45 (66%)**	0.32 (48%)	**0.39 (66%)**	0.28 (47%)	**0.69 (77%)**	0.62 (69%)
Malignant	1 Point	0.28 (48%)	0.25 (42%)	0.21 (45%)	0.18 (40%)	0.57 (70%)	**0.49 (60%)**

binary classifier performs better than ResNet18 for benign class, while ResNet18 performs better for malignant class.

Table 8: Comparison of segmentation performance on held out validation set for ResNet18 3-class classifier and ResNet34 2-class classifier using GradCAM

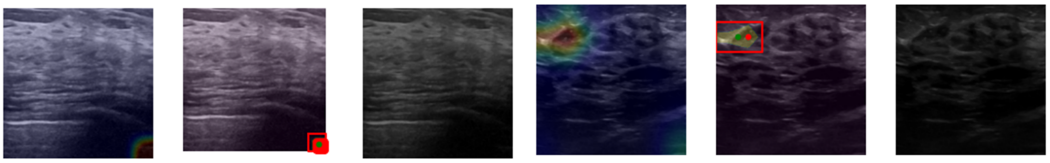
Type of Tumor	Prompt	Avg Dice Score		Avg IOU		Avg Pixel Accuracy	
		ResNet18(3-class)	ResNet34(2-class)	ResNet18(3-class)	ResNet34(2-class)	ResNet18(3-class)	ResNet34(2-class)
Benign	Box	0.37 (43%)	**0.47 (54%)**	0.30 (39%)	**0.39 (50%)**	0.94 (95%)	0.94 (96%)
Malignant	Box	**0.32 (39%)**	0.17 (20%)	**0.23 (33%)**	0.12 (17%)	0.87 (91%)	0.83 (87%)
Benign	Box + 1 Point	0.33(38%)	**0.46 (53%)**	0.27 (34%)	**0.38 (49%)**	0.93 (94%)	0.94 (96%)
Malignant	Box + 1 Point	**0.33(40%)**	0.16 (20%)	**0.24 (34%)**	0.12 (16%)	0.87 (91%)	0.83 (87%)
Benign	1 Point	0.40 (59%)	**0.45 (66%)**	0.35 (59%)	**0.39 (66%)**	0.67 (75%)	0.69 (77%)
Malignant	1 Point	**0.35 (59%)**	0.28 (48%)	**0.27 (59%)**	0.21 (45%)	0.65 (65%)	0.57 (70%)

4.4 Visualizing prompts and CAMs:



(a) Prediction: 1 GT: 1

(b) Prediction: 0 GT: 1



(c) Prediction: 0 GT: 0

(d) Prediction: 1 GT: 0

Figure 3: ResNet 34 GradCAM with in each set of 3 figures. Left figure as CAM, middle as Psuedo mask with prompts and right as ground truth mask, 0 : normal 1: abnormal

5 Conclusion:

We were able to achieve DICE 0.45 for benign and 0.39 for malignant class which is 65% and 60% of Breast-SAM methodology Also we observed more than 5% increase in IOU and DICE when center point is used is used over max-point as point prompt We observed that smaller threshold (bigger box) leads drop in performance. Also, affects location of center point of mask. Also in nearly all cases prompts generated using pseudo masks of GradCAM performed better across all metrics than those generated using GradCAM++.ResNet18 (3 class classification) generated prompts work well for Benign, ResNet34 (binary classification) generated prompts work well for Malignant.

References

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. ISSN 2352-3409. doi: https://doi.org/10.1016/j.dib.2019.104863. URL https://www.sciencedirect.com/science/article/pii/S2352340919312181.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. doi: 10.1109/WACV.2018.00097.

Can Cui, Ruining Deng, Quan Liu, Tianyuan Yao, Shunxing Bao, Lucas W. Remedios, Bennett A. Landman, Yucheng Tang, and Yuankai Huo. All-in-sam: from weak annotation to pixel-wise nuclei segmentation with prompt-based finetuning. *Journal of Physics: Conference Series*, 2722 (1):012012, mar 2024. doi: 10.1088/1742-6596/2722/1/012012. URL https://dx.doi.org/10.1088/1742-6596/2722/1/012012.

François-Guillaume Fernandez. Torchcam: class activation explorer. https://github.com/frgfm/torch-cam, March 2020.

Mingzhe Hu, Yuheng Li, and Xiaofeng Yang. Breastsam: A study of segment anything model for breast tumor detection in ultrasound images, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1), January 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-44824-z. URL http://dx.doi.org/10.1038/s41467-024-44824-z.

Saiyang Na, Yuzhi Guo, Feng Jiang, Hehuan Ma, and Junzhou Huang. Segment any cell: A sam-based auto-prompting fine-tuning framework for nuclei segmentation, 2024.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Qi Wu, Yuyao Zhang, and Marawan Elbatel. Self-prompting large vision models for few-shot medical image segmentation, 2023.