Accelerating Medical Image Segmentation with EfficientViT-SAM

Isaac (Zack) Duitz
Department of Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139
duitz@mit.edu

Sophie Guo
Department of Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139
sophiejg@mit.edu

Sam Mitchell
Department of Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139
sammit@mit.edu

Evan Rubel
Department of Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139
erubel@mit.edu

Collin Wen
Department of Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139
wenc@mit.edu

Abstract

Advances in computer vision have opened new doors in medical imaging, where accurate and efficient segmentation of organs and abnormalities is essential for diagnostics and treatment planning. In this project, we apply the EfficientViT-SAM model, an optimized version of Meta's Segment Anything Model (SAM), to segment medical images with a focus on balancing precision and processing speed. EfficientViT-SAM combines SAM's prompt-based segmentation abilities with a streamlined, lightweight design, making it well-suited for handling large medical datasets that require rapid analysis. We begin by testing its zero-shot segmentation performance on a large diverse dataset of medical images. We also evaluate the performance of Medficient-SAM and the performance of EfficientViT-SAM after finetuning on medical images. Through our study, we aim to demonstrate how this approach can support clinicians in their workflows, reducing workload and enhancing accuracy in medical image processing. We hope this project offers practical insights into using efficient segmentation tools to improve patient care, especially in settings where quick, reliable results are crucial.

1 Introduction

Computer vision has achieved remarkable advancements in medical imaging, where precise and efficient image segmentation is critical for clinical applications. Image segmentation involves partitioning an image to isolate regions of interest, such as organs or abnormalities, which supports diagnosis and treatment planning.

This project focuses on applying the Efficient Segment Anything Model (EfficientViT-SAM) to segment medical images. EfficientViT-SAM integrates the Segment Anything Model (SAM)—a prompt-able segmentation system known for its zero-shot generalization capabilities—with a lightweight

architecture designed to improve segmentation efficiency [1]. Efficient segmentation is especially valuable in medical imaging, where the need for rapid and accurate processing of large datasets is essential for timely diagnosis.

In this study, we evaluate the segmentation performance of three models: **EfficientViT-SAM**, **Medficient-SAM**, and a fine-tuned variant called **EfficientViT-MedSAM**. Using bounding boxes as prompts, we compare their accuracy to identify optimal solutions for improving efficiency and accuracy in medical image segmentation. EfficientViT-SAM serves as a general-purpose baseline, Medficient-SAM as an industry standard [2], and EfficientViT-MedSAM as an intermediate approach that bridges the gap between these two models. Unlike Medficient-SAM, which is fully optimized for medical imaging tasks, EfficientViT-MedSAM lacks specialized domain knowledge but still leverages the lightweight architecture of EfficientViT-SAM for improved efficiency.

We assess these models on the CVPR 2024 validation dataset, which includes a diverse set of modalities and medical imaging types, such as 2D and 3D scans from Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and others.

By analyzing the performance across these modalities, we aim to evaluate the trade-offs between general-purpose models and domain-specific optimizations. This investigation highlights the potential of intermediate approaches, such as EfficientViT-MedSAM, for balancing segmentation efficiency and accuracy in medical imaging without relying on extensive domain-specific adaptations.

2 Related Work

2.1 2D and 3D Segmentation

There is substantial research interest in segmenting objects in both 2D and 3D. For 2D images, state-of-the-art methods include Mask R-CNN [3] (which also performs object detection upstream) and 2D UNets [4], and both methodologies are based on convolutional networks. For 3D images and videos, 3D UNets [5] are the standard method for generating these higher-dimensional segmentations.

2.2 Promptable Visual Segmentation

Since the development of Mask R-CNN and UNets, the field has shifted to the task of **Promptable Visual Segmentation**, where the objective is to obtain a segmentation mask for a given image or video along with a prompt to help the model localize the object of interest. The prompt may be a bounding box, positive or negative points, free-form scribbles, or even text. Meta's Segment Anything Model (SAM) yielded remarkable results in zero-shot PVS for 2D images [6]; its extension for video — SAM 2 — leverages a memory module to track objects between adjacent frames, even when those objects become occluded or otherwise disappear between frames [7].

2.3 Segmentation for Medical Images and Video

Segmentation is a crucial component of the clinical workflow, as it is used in diagnostics and treatment planning. However, it is laborious and often requires specialist knowledge — automated methods, therefore, offer a clear benefit for clinicians and their patients. Several of the aforementioned approaches have been applied to medical images and video, especially for the task of PVS. SAM and SAM2 do not perform well in zero-shot evaluations with medical imaging because they are mostly trained on natural images. To mitigate this issue, MedSAM [8] and MedSAM-2 [9] are fine-tuned on more diverse medical image and video datasets. Moreover, nnUNet offers performant biomedical image segmentation in both 2D and 3D, automatically adapting to the input dataset and yielding the optimal network configuration and hyperparameters [10].

2.4 Efficient Segmentation

Cutting-edge PVS models (e.g., SAM and SAM2) are quite large and inefficient; with their important downstream uses, there is significant interest in making these models more efficient. Most recently, EfficientViT-SAM leverages knowledge distillation techniques – alongside pruning, quantization, and hardware-aware neural architecture search – to achieve a 48.9x speedup over SAM-ViT-Huge without a degradation of accuracy [1].

3 Method

3.1 Task

In this project, we undertake two primary tasks to evaluate the performance of EfficientViT-SAM, Medficient-SAM, and EfficientViT-MedSAM. First, we fine-tune EfficientViT-SAM on the CVPR training dataset to create EfficientViT-MedSAM, adapting it specifically for segmentation tasks in the medical domain. Second, we perform testing on all three models using the CVPR evaluation dataset to assess their performance without additional fine-tuning. During evaluation, we use the ground truth bounding box as prompts for all models to guide the segmentation process. For evaluation metrics, we compute the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) across all models to compare their ability to segment medical images effectively. Additionally, we measure the throughput time for each model to assess their efficiency in processing large datasets. This dual evaluation of accuracy and efficiency provides a comprehensive analysis of the trade-offs between general-purpose and specialized segmentation models in medical imaging.

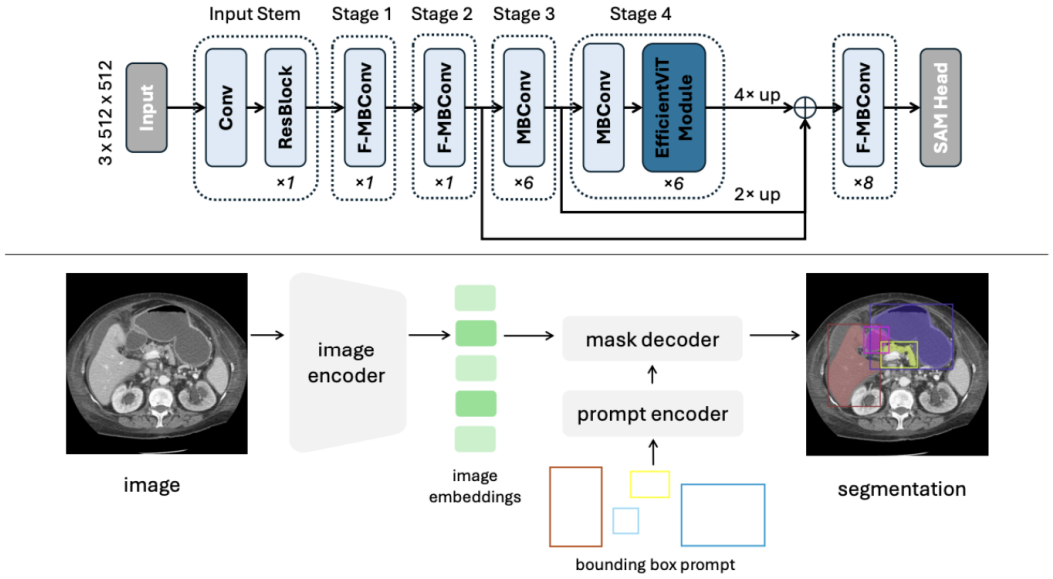

Figure 1: EfficientViT-SAM's macro-architecture of the image encode

3.2 Models

- EfficientViT-SAM: This model uses the EfficientViT architecture, where the image encoder is directly trained for the task of segmenting natural images. It leverages the Segment Anything Model (SAM) framework for promptable segmentation, enabling it to generalize to unseen objects with zero-shot capabilities. However, it lacks any specific optimization for medical image segmentation.
- Medficient-SAM: This model builds upon MedSAM, a domain-specific segmentation model, by employing knowledge distillation. The process transfers the representation capabilities of MedSAM's image encoder (a computationally intensive ViT-B model) into the lightweight EfficientViT architecture. Following the knowledge distillation, the model undergoes end-to-end fine-tuning on medical images, balancing segmentation accuracy with computational efficiency for the medical domain.
- EfficientViT-MedSAM: This model begins with the EfficientViT-SAM and directly fine-tunes it using medical imaging data. Unlike Medficient-SAM, it does not use knowledge distillation. Instead, it focuses solely on adapting the general-purpose EfficientViT-SAM for medical segmentation tasks through task-specific fine-tuning.

3.3 Data

We validate our three models using the validation split of the CVPR dataset. The CVPR dataset contains 2D and 3D medical images from 11 different modalities to capture the wide variety of medical segmentation use cases. These include Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), X-ray, ultrasound, mammography, optical coherence tomography (OCT), endoscopy, fundus, dermoscopy, and microscopy. For each of these images, a ground truth pixel-wise segmentation is provided, as well as the bounded box enclosing the full range of the segmentation. Bounding boxes are 2D and 3D, respectively, for 2D and 3D images. For standardization, these images were all resized to be (1024, 1024, n) for segmentation.

For 2D segmentations, each segmentation mask will directly correspond to a ground truth bounding box. For 3D segmentations, each 2D slice will is segmented individually with the 3D bounding box, given that the z-coordinate of the slide falls in between at least one bounding box. In the event that there are no bounding boxes at a given height, no inference is performed.

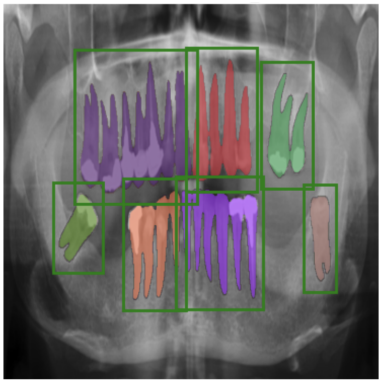

Figure 2: An instance from the CVPR dataset – a 2D X-ray of teeth with corresponding bounding boxes and segmentation masks

Table 1: Number of Images Across Dataset Modalities

Modality	Train Set	Validation Set
CT	14622	1140
Dermoscopy	3694	66
Endoscopy	43443	200
Fundus	1057	10
Mammography	1233	–
Microscopy	1000	50
MR	5144	628
OCT	1436	–
PET	546	3
US	1646	600
X-Ray	34893	379
Total	121718	3076

3.4 Metrics

In a medical setting, our measurement for accuracy will be through the area under the precision-recall curve. For a task such as tumor detection, this method ensures minimizing both false negatives and maximizing true positives. For measuring the accuracy of the segmentation itself, we will use the Dice similarity coefficient (DSC), which measures the overlap between the segmentation and ground truth. This metric accounts true positives, false positives, and false negatives into the score.

4 Results

From our results, we see that Medficient-SAM, with its knowledge distillation strategy, performs the best as exemplified by its high 3D dice score. Fine-tuning EfficientViT does indeed improve its accuracy, earning higher 2D and 3D dice scores.

Video Demo on CPU

Table 2: CVPR Validation Set Results Across Models (A100)

Model	2D Dice (%)	2D Inference Throughput (imgs/s)	3D Dice (%)	3D Inference Throughput (imgs/s)
EfficientVIT-SAM	57.39	162.76	29.57	3.92
EfficientVIT-MedSAM	81.21	160.40	41.86	4.61
Medficient-SAM	86.42	50	70.21	2.33

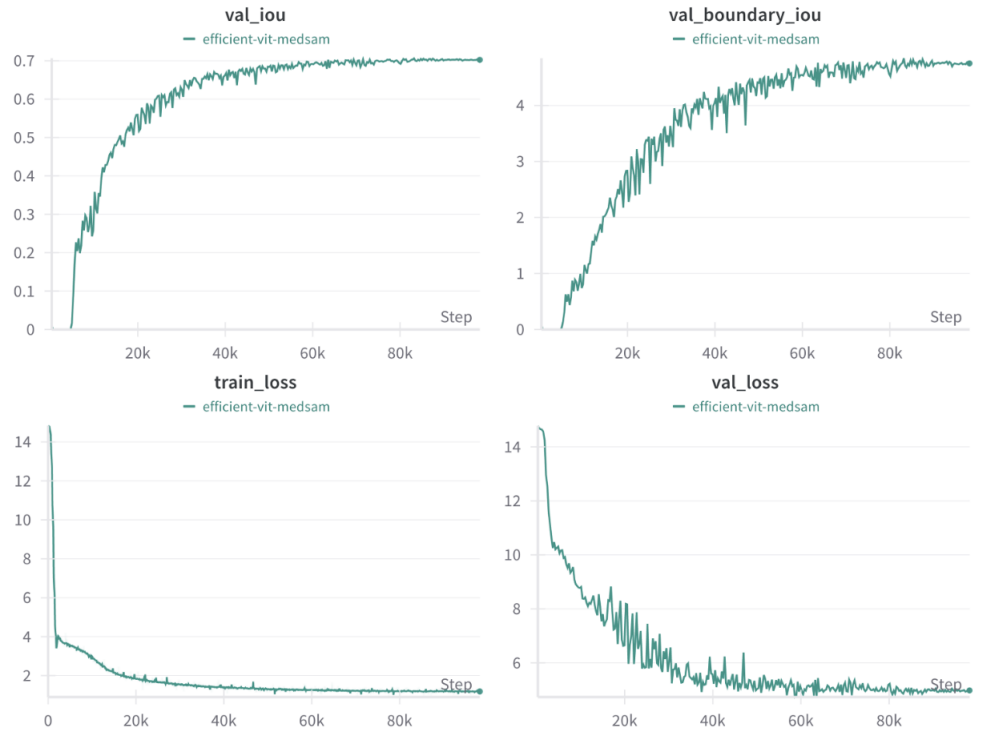

Figure 3: Training curves for finetuning EfficientViT-MedSAM

With EffcientViT-SAM and EfficientViT-MedSAM, we see similar performance metrics at about 160 2D images per second and about 4 3D images per second. We justify these results because these models have similar architectures based on the EfficientViT backbone. We also observe that, with the very diverse validation set, certain modalities (e.g., ultrasound and X-ray) are more difficult to segment than others, perhaps due to the unclear boundaries.

5 Conclusion

5.1 Discussion

In our study, we provide a thorough benchmark of three efficient segmentation models by validating on a diverse set of medical images.Here, we also release our full model – training code and weights – to inspire future improvements.

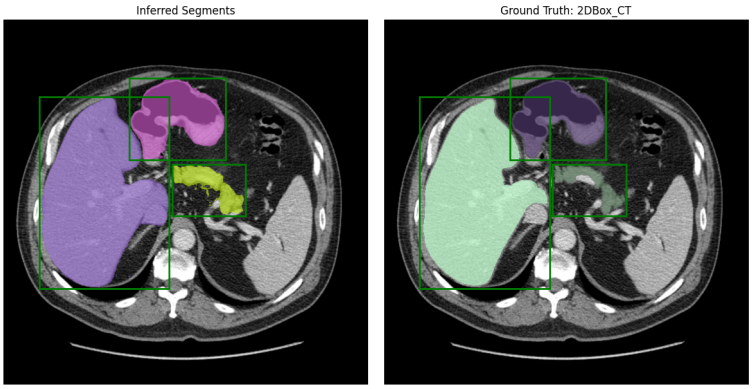

Figure 4: Inference example on 2D CT scan segmentation with EfficientVit-SAM

5.2 Future Work

There are many fascinating future directions for our work. Firstly, with patient privacy being paramount, we propose a federated learning scheme to train efficient medical image segmentation models, using private data from different institutions. We can also improve the efficiency of 3D-native segmentation models (such as SAM2), which use a memory module to propagate segmentations between frames and give more coherence to the 3D masks. Finally, we can explore ways to integrate our models into clinical workflows, especially in resource-constrained settings.

5.3 Group Member Contributions

All members participated equally in the ideation of this project. Isaac and Collin worked on the evaluation of Efficient-ViT-SAM and Medficient-SAM. Sam worked on the data aggregation and preprocessing steps. Sophie and Evan worked on the Efficient-ViT-SAM inference and training.

References

[1] Zhang, Zhuoyang, Han Cai, and Song Han. “Efficientvit-sam: Accelerated segment anything model without performance loss.” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2024.

[2] Le, Bao-Hiep, et al. "MedficientSAM: A robust medical segmentation model with optimized inference pipeline for limited clinical settings." CVPR 2024: Segment Anything In Medical Images On Laptop. 2024.

[3] He, Kaiming, et al. “Mask r-cnn.” *Proceedings of the IEEE international conference on computer vision.* 2017.

[4] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation.” *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18.* Springer International Publishing, 2015.

[5] Çiçek, Özgün, et al. “3D U-Net: learning dense volumetric segmentation from sparse annotation.” *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19.* Springer International Publishing, 2016.

[6] Kirillov, Alexander, et al. “Segment anything.” *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 2023.

[7] Ravi, Nikhila, et al. “Sam 2: Segment anything in images and videos.” *arXiv preprint arXiv:2408.00714* (2024).

[8] Ma, Jun, et al. “Segment anything in medical images.” *Nature Communications* 15.1 (2024): 654.

[9] Zhu, Jiayuan, Yunli Qi, and Junde Wu. "Medical sam 2: Segment medical images as video via segment anything model 2." *arXiv preprint arXiv:2408.00874* (2024).

[10] Isensee, Fabian, et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation." *Nature methods* 18.2 (2021): 203-211.