Efficient Segment Anything Model for Medical Image Analysis

Yurvan Ramjan (221004686)
Supervisor: Prof. Serestina Viriri

***Abstract*—Medical image analysis is essential for accurate diagnosis and treatment planning but remains challenging due to the complex nature of anatomical structures and object boundaries. While the Segment Anything Model (SAM) has shown promise for general-purpose segmentation[1], its high computational demands limit its use in real-time clinical settings. However, medical image segmentation is more challenging due to complex modalities, fine anatomical structures, uncertain and complex object boundaries, and a wide range of object scales.**

An Efficient Segment Anything Model (EfficientSAM) is implemented in this research work, an optimized version of SAM, designed to improve segmentation performance across various medical imaging modalities (e.g., CT, MRI) while reducing computational complexity. Techniques such as Mixed-Precision Training and Batch Processing are employed to enhance efficiency. Results show that EfficientSAM achieves comparable segmentation accuracy to SAM while significantly reducing model size, inference time, and computational cost, making it more practical for clinical workflows.

The results achieved reduces the original model's parameter count, while maintaining segmentation performance and improvement on inference speed. EfficientSAM contains 86M parameters, a 86.4% reduction from the original 632M parameters. Results had shown a 3.55x increase in DICE score, and a 3.40x increase in Jaccard score. The code is available at EfficientSAM for Medical Image Analysis.

***Index Terms*—Segmentation, Segment Anything Model, Efficient Segment Anything Model, Vision Transformer, Medical Image Analysis**

I. Introduction and Background

Segment Anything Model (SAM) is an advanced image processing tool built on the Vision Transformer (ViT) architecture [1], trained on over eleven million images and one billion masks. This extensive training allows SAM to perform zero-shot segmentation on diverse datasets and tasks using points and bounding boxes for high-precision pixel-level semantics and regional contexts. However, while SAM excels in non-medical segmentation [2], its performance declines with medical images, especially for amorphous lesions, where precision is critical for clinical outcomes.

To address these challenges, this study introduces EfficientSAM, a lightweight variant of SAM tailored for medical imaging with similar accuracy and computational efficiency and propose the framework visualized in Figure 2, as well as compare EfficientSAM's architecture to an existing base vision transformer (ViT-B) of SAM. EfficientSAM retains SAM's generalization capabilities while optimizing performance on complex medical datasets, yielding both quantitative and qualitative improvements across modalities like CT, MRI, and Ultrasound.

SAM draws on pre-trained computer vision models such as Contrastive Language-Image Pre-training (CLIP) [3] and A Large-scale ImaGe and Noisy-text embedding (ALIGN) [4]. These models enhance its ability to perform zero-shot tasks, such as object detection and image classification [3], and generate natural language descriptions from images. Similarly, ALIGN enhances SAM's ability to generate natural language descriptions of image regions, providing detailed results beyond traditional captioning methods [4].

However, despite their impressive capabilities, these models (including DALL·E - an advanced AI model, by OpenAI, designed for generating images from textual descriptions [5]) remain underexplored in medical image segmentation.

This research aims to extend SAM's healthcare domain capabilities and introduce novel optimizations through EfficientSAM.

A. Segment Anything Model for Non-Medical Image Applications

SAM has shown mixed success in various non-medical applications. For example, Tang et al. [6] evaluated SAM on a challenging task - camouflaged object segmentation - which involved animals blending into their natural environments. Their study revealed that SAM struggled with accuracy under these conditions, suggesting that additional human guidance or prompt input may be necessary.

On the other hand, Ji et al. [7] embarked on an extensive exploration of SAM, investigating three testing methods (point, box, everything) across a diverse range of applications. Their research spanned natural imagery, agriculture, manufacturing, and remote sensing, uncovering SAM's strengths in segmenting objects in clear conditions but also exposed limitations in complex scenarios. Notably, the study emphasized the importance of human-provided prompts (e.g., points or boxes) to improve segmentation outcomes.

B. Segment Anything Model for Medical Image Analysis

In the medical domain, SAM's performance has been evaluated across various tasks, revealing both its potential and limitations. Ji et al. [8] tested SAM's segmentation capabilities

[1]Segmentation refers to the process of partitioning an image into distinct regions or segments, which correspond to different objects or parts of the scene.

on brain, lung, and liver lesions using CT and MRI scans. While SAM accurately segmented organs with well-defined boundaries, it struggled with amorphous lesion regions. Ji et al. [7] showed SAM's capabilities across several healthcare subfields, including optical disc and cup, polyp[2], and skin lesion segmentation. SAM performed well with substantial human prior knowledge (points or bounding boxes), but the absence of prompts led to inaccurate segmentations.

Mohapatra et al. [9] compared SAM to the traditional Brain Extraction Tool (BET) in neuroimaging. BET, although widely used, suffers from limitations such as over-extraction in lesioned brains and susceptibility to poor image quality [9]. SAM outperformed BET across several metrics, including Dice coefficient, Intersection over Union (IoU), and other accuracy metrics, especially when images contained brain lesions near the edges or meninges[3].

Other studies have explored SAM's use in digital pathology. Deng et al. [10] evaluated SAM's segmentation of tumor regions and cell nuclei in whole-slide imaging. While SAM excelled at segmenting large connected objects, it was inconsistent with dense instance segmentation, even when provided with multiple prompt points per image. Similarily, Zhou et al. [11] applied SAM to polyp segmentation using five benchmark datasets, noting that although SAM performed well in some cases, it fell short of state-of-the-art (SOTA) methods. Liu et al. [12] integrated SAM with a 3D Slicer software tool [13] to facilitate and unlock a new dimension for medical image analysis, further highlighting the model's potential but also pointing out the need for fine-tuning.

Several studies have confirmed SAM's moderate performance across diverse Medical Image Segmentation (MIS) datasets [14], [15], [16], [17]. For instance, He et al. [14] reported that SAM's zero-shot performance lagged behind conventional deep learning methods, while Mazurowski et al. [15] observed that SAM's performance converges as the number of points increase and this performance varied significantly across datasets. These findings underscore the importance of fine-tuning and optimization for SAM to become practical for clinical use.

To address the limitations of SAM, researchers have developed specialized models. MedSAM, proposed by Ma et al. [16], achieved a remarkable 22.51% increase in DICE score compared to SAM, but at the cost of increased computational demands. Similarly, Wu et al. [17] applied adapter-based fine-tuning to improve SAM's performance for medical applications, but the model's size and complexity still pose challenges. These efforts highlight the need for a lightweight, efficient model capable of balancing accuracy and computational efficiency - a gap this study aims to fill through the development of EfficientSAM.

[2]A polyp is an abnormal growth of tissue that arises from a mucous membrane and projects into a hollow organ, such as the stomach or the intestine.

[3]The three layers of membranes that protect the brain and spinal cord

The problem addressed in this study is: How can SAM be optimized to improve segmentation performance for medical images while reducing computational complexity to enable medical image analysis?

This research paper unveils the meticulous construction of a vast, custom-built medical image dataset, featuring 10 diverse image modalities (refer to Figure 1), offering comprehensive coverage of the human body. This dataset served as a vital tool for thoroughly analyzing and evaluating EfficientSAM's performance in the realm of medical image segmentation.

Moreover, the paper showcases compelling quantitative and qualitative experimental results, providing an in-depth evaluation of the correlation between EfficientSAM's performance and the dataset. This study used the following enhancements to achieve EfficientSAM's objective:

- Optimize SAM's architecture to reduce computational overhead using a lighter, pre-trained ViT backbone and fine-tuning the weights, which results in parameter reduction
- Reduce inference time by utilizing Mixed-Precision training.
- Evaluate EfficientSAM on multiple modalities (CT, MRI).
- Compare performance with other models using Dice coefficient and IoU metrics against EfficientSAM.

This is an effective solution that bridges the gap between performance and efficiency in clinical settings.

C. Structure of the Paper

This paper is organized as follows: Section II discusses the Literature Review, Section III details the Methods and Techniques, Section IV presents the Experimental Results and Discussion as well as analyzes the findings, and Section V concludes the study.

II. Literature Review

A. Segment Anything Model in Real-World Applications

Ji et al. [7] discuss several aspects of SAM's performance across different scenarios. The authors claim that SAM demonstrates exceptional adaptability in ordinary scenes and proves to be effective across various prompt modes, especially when targeting regions that are distinct from their surroundings. Nonetheless, in more complex scenes such as fundus[4] image segmentation, SAM may require additional manual prompts with prior knowledge, potentially leading to suboptimal user experiences [7]. It tends to favour selecting the foreground mask, resulting in poor performance in tasks like shadow detection due to foreground bias in its pre-training dataset [7].

The article additionally claims that SAM struggles with low-contrast applications and faces challenges in seamlessly segmenting objects embedded in their surroundings. In real-world medical applications, SAM tends to yield unsatisfactory

[4]Fundus refers to the network of blood vessels located in the interior surface of the eye, opposite the lens, including structures such as the retina, optic disc, etc.

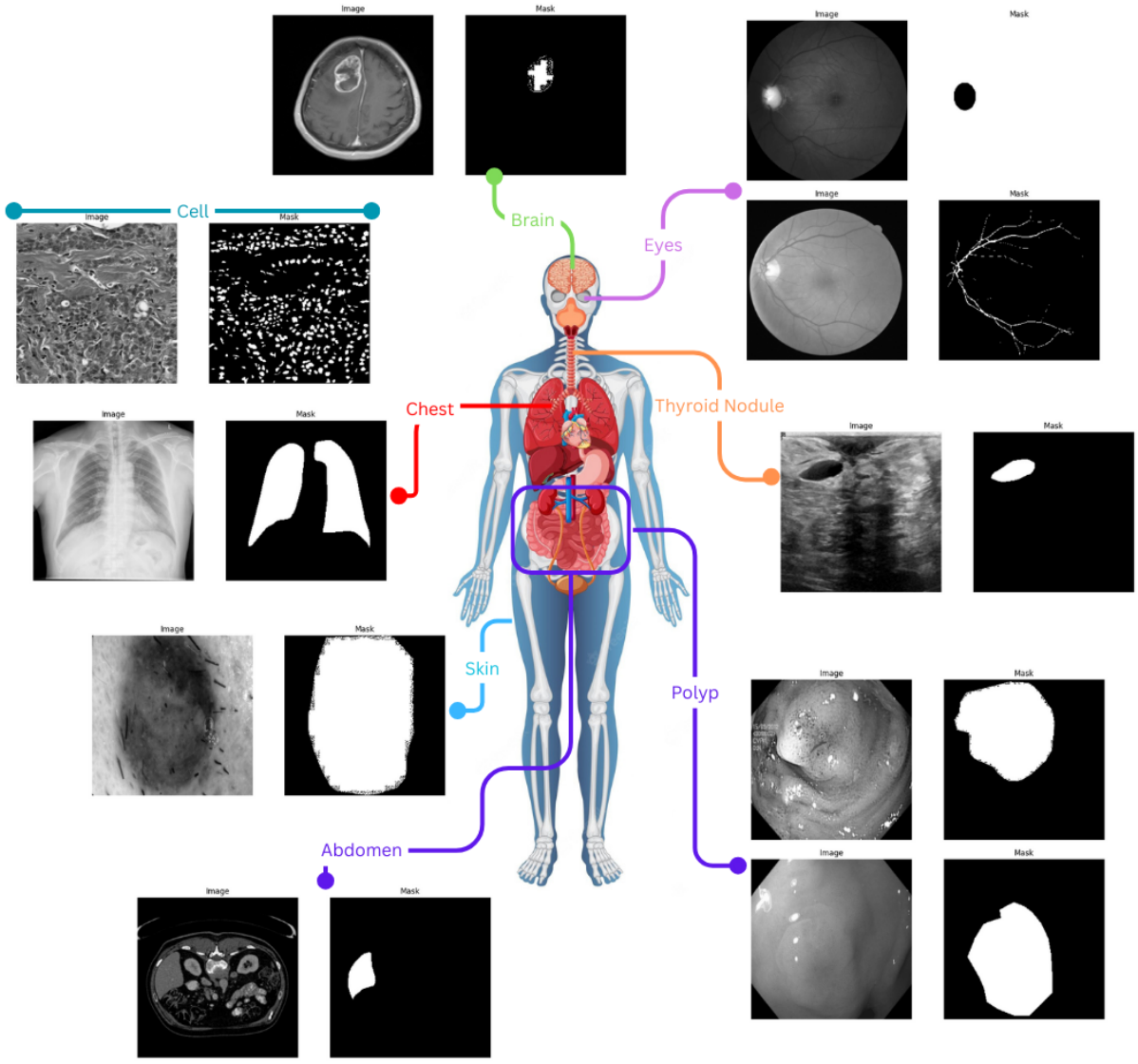

Fig. 1: The combined dataset showing the images and binary mask(s) that cover the majority of biomedical objects, for example, brain tumours, fundus vasculature, thyroid nodules, lung, abdominal organs and tumors, cell, and polyp.

results, particularly in box and everything modes, highlighting its limitations in comprehending practical scenarios [7]. While SAM performs well in standard scenarios, its limitations in complex tasks such as fundus image segmentation and shadow detection highlight the need for models with enhanced adaptability and domain-specific tuning. This motivates the development of EfficientSAM to address these limitations by reducing dependency on manual prompts and overcoming foreground bias.

B. Segment Anything Model in Medical Imaging

Ma et al. [16] demonstrate SAM's exceptional ability to generalize, surpassing other deep learning-based interactive segmentation methods that are often limited in scope. The article highlights SAM's extensive evaluation across various medical images, delivering acceptable segmentation results, especially for targets with clear boundaries. To combat this, the authors introduced MedSAM as an enhanced model, improving SAM's segmentation performance on medical images. Its capabilities have the prospective to accelerate the advancement of diagnostic tools and personalized treatment procedures, demonstrating MedSAM's remarkable generalization ability and its effectiveness in addressing SAM's segmentation challenges [16].

Although MedSAM improves upon SAM's performance for medical images, challenges such as faint boundaries and reduced contrast persist [16]. This suggests an unmet need for further innovation, motivating the exploration of EfficientSAM as a lightweight, adaptable solution tailored for high-resolution

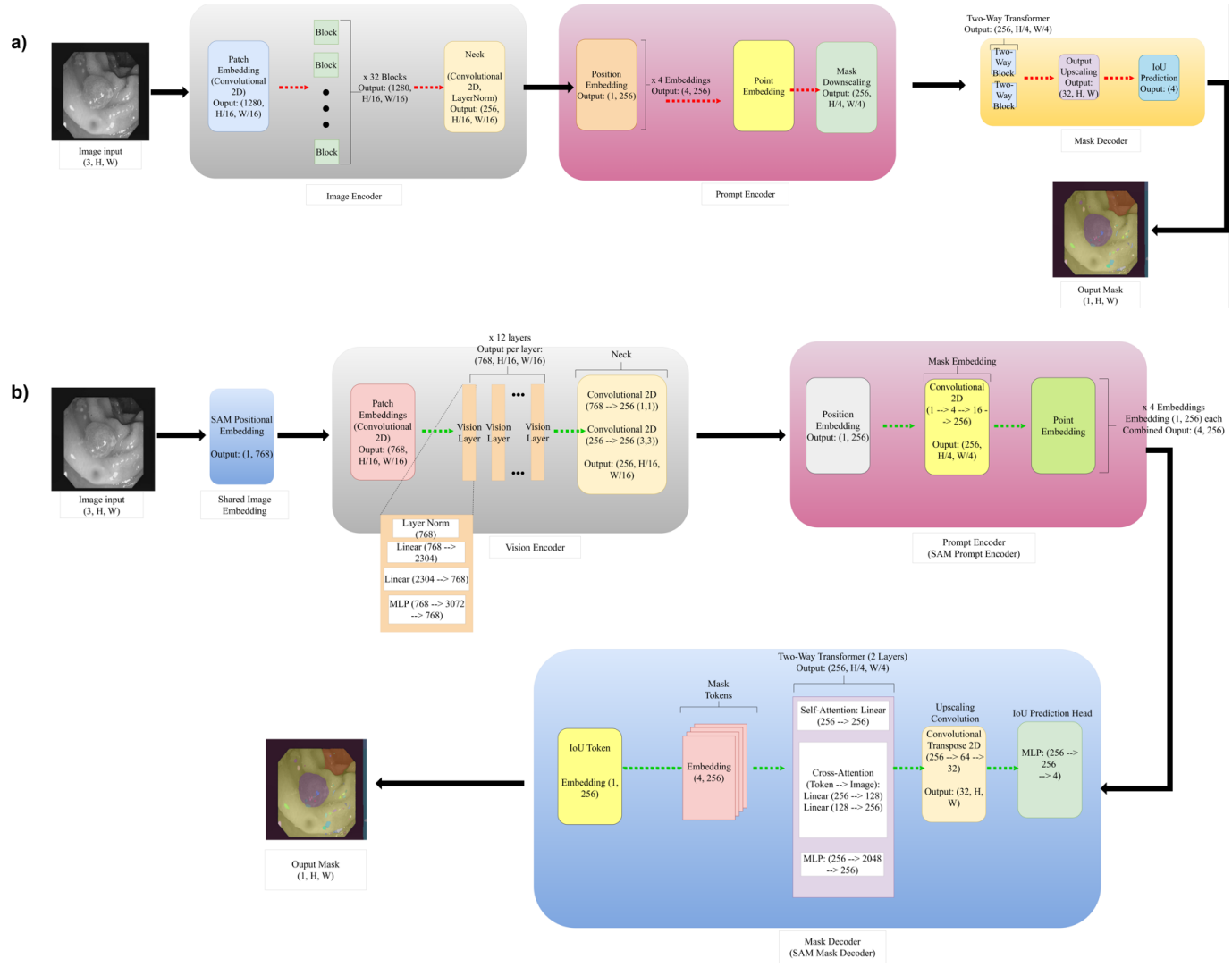

Fig. 2: Comparison of the Segment Anything Model Architectures. **(a)** Overview of the Base Vision Transformer Architecture for Segment Anything Model. **(b)** Overview of the Proposed Architecture for Efficient Segment Anything Model (EfficientSAM).

medical segmentation tasks.

C. Accelerating Segment Anything Model Without Performance Loss

Zhang et al. [18] introduced a groundbreaking solution, EfficientViT-SAM, to accelerate the capabilities of the Segment Anything Model (SAM). By seamlessly integrating the efficient and powerful EfficientViT (Efficient Vision Transformer) as the image encoder while maintaining SAM's lightweight prompt encoder and mask decoder, this innovation achieves an exceptional 48.9x measured Tensor Runtime (TensorRT)[5] speedup on A100 GPU over SAM-ViT-H [19], without compromising performance [18].

While EfficientViT-SAM achieves significant runtime improvements, its reliance on TensorRT and high-end hardware limits its applicability in resource-constrained environments. This creates an opportunity for lightweight models like EfficientSAM, which balance efficiency and performance for broader adoption in real-world clinical settings

As highlighted in the article, the quest to replace SAM's image encoder with lighter models like MobileSAM [20], EdgeSAM [21], [18], and EfficientSAM [18] has been ongoing. However, these alternatives often led to substantial performance drops. EfficientViT-SAM not only overcomes this obstacle but also sets the stage for more efficient and effective training. The two-phase training approach outlined in the article demonstrates a meticulous process that underlines the significance of this groundbreaking solution.

According to the article [18], the pursuit to accelerate SAM has witnessed various strategies, such as MobileSAM [20] distilling the knowledge of SAM's ViT-H model into a compact vision transformer. Similarly, EdgeSAM [18], [21] and EfficientSAM have also contributed to this evolution with their unique methodologies and approaches, along with the CNN-based architecture stated in the previous article. EfficientSAM leverages the Mask AutoEncoder (MAE) pretraining method

[5]TensorRT is a high-performance deep learning inference library developed by NVIDIA.

to improve performance.

MobileSAM sacrifices performance, whereas EfficientViT-SAM improves speed but increases complexity. EfficientViT-SAM is extensively evaluated on zero-shot benchmarks, demonstrating significant performance and efficiency enhancements compared to previous SAM models.

D. Pretraining Techniques for Efficient Segment Anything

Xiong et al. [22] compared EfficientSAM, MobileSAM, FastSAM, and SAM in terms of throughput, parameters, and performance for zero-shot instance segmentation in the Common Objects in Context (COCO) dataset. The article shows that the leveraging masked image pretraining method Segment Anything Model for Medical Images (SAMI), which reconstructs features from the SAM image encoder for visual representation learning, resulting in EfficientSAM with SAMI-pretrained lightweight image encoders perform favourably, outperforming other methods by approximately 4 Average Precision (AP) on COCO/LVIS (Common Objects in Context/Large Vocabulary Instance Segmentation).

The article [22] highlights the significant disparity in parameters between the ViT-Huge (ViT-H) image encoder and the prompt-based decoder in SAM. With 632 million parameters in the ViT-H encoder and just 3.87 million in the prompt-based decoder, the computation and memory costs associated with using SAM for segmenting tasks present a significant obstacle for real-time applications. However, recent research efforts have proposed innovative strategies to mitigate these challenges and minimize the computational burden when employing SAM for prompt-based instance segmentation. One compelling approach outlined in the article [22] involves distilling knowledge from the default ViT-H image encoder into a smaller ViT image encoder. By doing so, the computational cost can be substantially reduced, enabling the utilization of a real-time Convolutional Neural Network (CNN)-based architecture for the Segment Anything task.

Furthermore, the use of pre-trained lightweight ViT image encoders, such as ViT-Tiny or ViT-Small, to streamline SAM upholds the performance standards [22]. Notably, SAM-leveraged masked image pretraining (SAMI) leverages the Mask AutoEncoder (MAE) pretraining methodology in conjunction with the SAM model to attain high-quality pre-trained ViT encoders [22]. By leveraging a SAM encoder, ViT-H, to create feature embeddings and train a masked image model with lightweight encoders, SAMI achieves the reconstruction of features from ViT-H of SAM instead of image patches. This results in generalized ViT backbones that are well-suited for diverse tasks including image classification, object detection, and segmenting. Ultimately, pretrained lightweight encoders are fine-tuned with SAM decoders to fulfill the requirements of the segmenting task [22].

Although pretraining techniques like Mask AutoEncoder (MAE) improve segmentation, they are not sufficient for real-time medical applications due to the high memory footprint of models like ViT-H. EfficientSAM leverages lightweight encoders to address this gap, ensuring faster inference without sacrificing segmentation quality. This strategy gains support from SAMI, which delivers exceptional pre-trained ViT backbones for segmenting tasks.

E. Comparison of Vision Transformer (ViT) Architectures

Huang et al. [2] compared the performance between the ViT-H (Vision Transformer-High) and ViT-B (Vision Transformer-Base) architectures for evaluating SAM's effectiveness in medical image segmentation. The findings indicate that ViT-H outperforms ViT-B in terms of segmentation accuracy and robustness. Specifically, ViT-H exhibited greater stability and adaptability across various medical datasets, particularly for complex anatomical structures and challenging segmentation scenarios. The larger parameters and architecture of ViT-H allowed it to better capture subtle details and intricate features essential in medical contexts [2]. In contrast, the smaller ViT-B model demonstrated more inconsistency [2], particularly struggling with cases that involved complex object boundaries or limited contrast between regions.

According to Huang et al. [2], SAM with ViT-H demonstrates superior segmentation performance compared to ViT-B across a range of tasks, including fine-grained medical segmentation. However, even though the ViT-H achieves better results, it possesses higher memory consumption, longer inference times, and increased power requirements.

Notably, both models benefited from task-specific fine-tuning, with ViT-H achieving a more significant performance gain of approximately 6.68%, compared to ViT-B's improvement of 4.39% in DICE scores [2]. This suggests that while ViT-B can serve as a lightweight model for quicker tasks, ViT-H provides more reliable performance for demanding segmentation needs, especially when fine-tuned on medical datasets like COSMOS 1050K [2]. Although ViT-H shows superior performance in complex scenarios, both models are sensitive to input prompts and require careful handling to maintain consistent results.

The implementation of EfficientSAM in medical image analysis, the ViT-B architecture was selected over ViT-H due to several key advantages that align with the goal of developing a lightweight, efficient model:

- **Model Complexity and Efficiency:** ViT-H, with more parameters and higher computational demands, directly contradicted the efficiency goals of EfficientSAM. ViT-B, being smaller and faster, required fewer resources for training and deployment, which made it ideal for real-time applications in resource-constrained environments, such as edge devices or smaller medical institutions.
- **Reduced Inference Time:** Clinical settings often demand faster inference inference speeds to support quick decision-making. ViT-B significantly reduced the inference time compared to ViT-H, making it appropriate for devices with limited memory while ensuring efficient medical image analysis.
- **Diminishing Returns on Performance:** Although ViT-H offered higher accuracy in complex segmentation tasks, the performance gains were not always worth the increased resource consumption. For simpler segmentation tasks (e.g., polyp detection or lung segmentation), ViT-B

performed sufficiently well without unnecessary computational overhead.
- **Ease of Fine-Tuning:** Fine-tuning ViT-B was more efficient and straightforward, facilitating seamless deployment of EfficientSAM across the various datasets. Its smaller size also mitigated the risk of overfitting, which is a concern when working with limited annotated medical data.
- **Energy Efficiency and Scalability:** ViT-B consumed less computational resources, which makes it more scalable across various devices and infrastructures, aligning with the sustainability objectives of modern medical AI projects.

Choosing ViT-B ensured that the EfficientSAM implementation remains fast, lightweight, and adaptable, without sacrificing too much accuracy. This approach resulted in a practical, deployable solution that balanced performance with real-world constraints. The lower parameter count and reduced computational complexity of ViT-B made it easier to train, faster to execute, and better suited for efficient medical image analysis.

Despite ViT-H's superior accuracy, its high computational demand and memory consumption render it impractical for edge devices and small clinics. ViT-B offers a more practical alternative, but its inconsistency in complex cases leaves room for further exploration, justifying the focus on developing a robust, lightweight version with EfficientSAM.

F. Efficient Segment Anything Models

Shu et al.[23] redefine the segment anything task by transforming it into an instance segmentation task with only one foreground category by leveraging the You Only Look Once version 8 (YOLOv8), calling it the Fast Segment Anything Model (FastSAM). FastSAM's approach to promptable segmentation, combined with a post-process strategy, holds significant potential for efficient and accurate segmentation. Despite the reformulated framework not yet matching SAM's performance on downstream zero-shot tasks, it represents a crucial advancement in this domain.

Furthermore, the article underscores recent initiatives aimed at developing a more computation-efficient segment anything model. For example, the Mobile Segment Anything Model (MobileSAM) [20] endeavors to replace the heavy component of the image encoder with the lightweight architecture of Tiny Vision Transformer (TinyViT). Although MobileSAM's approach involves a decoupled knowledge distillation strategy leading to performance decay due to the absence of supervision for the final mask prediction [20], this work stands as a notable stride towards achieving more efficient segment anything models.

While MobileSAM and FastSAM represent important strides toward efficient segmentation, they do not fully address the computational and accuracy trade-offs needed for real-time medical applications. EfficientSAM aims to bridge this gap by leveraging novel architectures like ViT-B, optimized with MAE pretraining, to ensure both speed and performance.

G. Relevance of Mixed-Precision

Mixed precision techniques have transformed deep learning by enhancing computational efficiency during training and inference. Micikevicius et al. [24] showcased that mixed precision leverages a combination of 16-bit (FP16) and 32-bit (FP32) floating-point arithmetic to dramatically reduce memory usage and accelerate training without compromising model accuracy. The incorporation of loss scaling (a crucial method to prevent gradient underflows) ensured stable optimization, even when using lower-precision computations [24]. This made mixed precision the optimal choice for large-scale models, enabling practitioners to push the limits of neural networks and facilitate faster experimentation and deployment.

Hardware-optimized mixed precision approaches [25] leverage computational memory architectures like Phase-Change Memory (PCM)[6], enabling complex model execution on edge devices [21], [18]. This is particularly important in healthcare for real-time diagnostics and monitoring.

Dörrich et al. [26] highlighted significant improvements in training speed and inference efficiency with mixed-precision across GPUs (Graphics Processing Units), TPUs (Tensor Processing Units), and AI-specific accelerators, emphasizing the need to balance precision and performance to minimize power consumption. This balance is essential for green AI[7] and sustainable computing.

Mixed precision is critical for training advanced models like the Segment Anything Model (SAM) and its lightweight variant EfficientSAM, which utilizes FP16 to optimize memory and training time, crucial for processing high-resolution medical images efficiently.

In summary, mixed-precision is not just a software optimization; it is vital for scalable, sustainable AI, especially in medical image segmentation tasks. It ensures efficient training and deployment of neural networks while balancing speed, accuracy, and power consumption, ultimately improving patient outcomes and reducing operational costs in healthcare settings.

III. Methods and Techniques

A. Introduction to Existing Segment Anything Models

Building upon the limitations identified in existing Segment Anything Models (SAM) for medical image analysis, EfficientSAM aims to tackle the following gaps:

1) **Limited Datasets**. Previous studies [9], [10], [11], [8], [7] have primarily evaluated SAM's performance in modalities like MRI, CT, and digital pathology using a minute number of segmented objects. However, this hindered comprehensive analysis in the field of medical image segmentation due to the vast image modalities and various anatomical structures that also require segmentation.

[6]PCM can store intermediate computations and models, speeding up both training and inference by minimizing the reliance on slower, power-hungry DRAM or external storage

[7]Green AI refers to the development and use of energy-efficient and environmentally sustainable artificial intelligence methods, promoting efficient algorithms, hardware, and practices to reduce the environmental footprint of AI while maintaining or even improving performance.

2) **Single Testing Strategy**. Most previous studies [8], [11], [15] evaluated SAM using a limited or single testing strategy, even though different medical objects may require different testing modes due to their unique characteristics, hence resulting to an inaccurate and incomplete analysis.
3) **Inadequate Assessments**. Some studies [7] only assessed SAM using visualization results from SAM's online demo or focused solely on limited metrics [10] like DICE or IoU. This has led to a lack of comprehensive analysis regarding SAM's perception of medical objects and its segmentation performance [14], [15], [16].

By designing a more efficient evaluation framework, testing across multiple datasets, and optimizing SAM's runtime for different medical modalities, we can gain a thorough understanding of EfficientSAM's segmentation performance and its perception of medical objects.

B. Dataset Overview

Medical imaging encompasses a variety of modalities such as CT, MRI, and X-ray, each with its unique strengths in visualizing anatomical structures and lesions. To thoroughly evaluate SAM's performance in medical image segmentation, 27 public datasets (see Table I) were meticulously curated and standardized to create a single expansive dataset covering these image modalities. This dataset's categorization system draws from the official introductions of each public dataset.

The details of the dataset are described in the following two aspects.

1) Dataset Collection: Table I displays a detailed list of the collected medical image datasets for segmentation.

2) Dataset Preprocessing Specification: The dataset encompasses a diverse range of data with various labels, formats, and shapes. The original SAM version solely supported 2D input, a crucial component for 3D/4D format.

For 3D volumes:

1) Extract slices along the main viewing plane for higher resolution.
2) Retain slices with a sum of pixel values of their labels greater than 50 for 3D image and label volumes.
3) Normalize the extracted image intensities using min-max normalization: $I_n = 255 * \frac{I - I_{min}}{I_{max} - I_{min}}$, limiting the range to (0, 255). I represents the original extracted image, I_n represents the normalized image, and I_{min} and I_{max} are the minimum and maximum intensity values of I respectively.
4) Save images and labels in PNG format.

For 4D data (N, W, H, D), convert the data into N sets of 3D volumes and then follow the 3D volume processing flow.

For 2D images:

1) Retain images with a sum of pixel values of their labels greater than 50.
2) Reset the pixel value of labels according to the object category or location within the range of 1 to 255.
3) Convert the format of images and labels from BMP, JPG, TIF, etc. to PNG for consistent data loading.

Preprocessing each of the collected datasets ensures that the segmentation model receives consistent input formats, standardizes the data across different datasets, which in turn enhances both performance and generalizability. Normalization ensures uniform intensity distribution across images, while filtering slices with low label pixel sums ensures that only meaningful data is used.

C. Investigating Inference Efficiency

The assertion on whether inference efficiency could be improved through mixed-precision training and batch processing was validated through a series of experiments. These optimizations demonstrated a reduction in computational overhead and accelerated testing by a factor of 10.66.

Mixed-precision training leveraged both 16-bit and 32-bit floating-point precision, reducing memory consumption and speeding up operations on the GPU and TPU. By storing the weights and activations in lower precision where appropriate, EfficientSAM's model inference time decreased the original SAM significantly without sacrificing segmentation accuracy. This technique not only shortened training and inference time but also reduced memory bottlenecks, which was particularly beneficial when working with high-resolution medical images.

SAM's original code design required the encoding operation to be performed multiple times per image, leading to poor runtime performance, especially for high-resolution inputs. To address this, the embedding features for all input images were pre-computed and saved as intermediate files. This allowed embeddings to be reused across multiple testing strategies, avoiding redundant encoding operations and reducing computational overhead by nearly 10 times.

Batch processing further improved efficiency by enabling parallel computation on multiple images. Instead of processing one image at a time, input images were grouped into batches, allowing the GPU to utilize its resources more effectively. This approach minimized idle time during computation and significantly increased throughput, enhancing runtime efficiency.

Validation of these techniques were implemented on four different testing strategies, as depicted in Table II, which shows the runtime performance of various Vision Transformer (ViT) architectures. The experiments revealed that using mixed-precision training, batch processing, and embedding reuse together drastically reduced inference time, resulting in a 10.66x improvement in runtime efficiency.

D. Segmentation Evaluation

To rigorously assess the effectiveness of EfficientSAM in generating accurate segmentations, we developed a mask-matching mechanism. This approach ensures that only the most precise mask predictions are selected for evaluation, thereby enhancing segmentation reliability across multiple test modes. The key objective of this method is to identify the mask

Dataset Name	Description	Image Modality
CAMUS (Cardiac Acquisitions for Multi-structure Ultrasound Segmentation) Dataset (Human Heart Data)	Four-chamber and Apical two-chamber Heart Segmentation	Ultrasound
Kvasir Dataset for Classification and Segmentation (Multi-Class Image-Dataset for Computer-Aided Gastrointestinal Disease Detection)	Multi-class Classification and Segmentation of Gastrointestinal Diseases	Colonoscopy
IXI Dataset	Structural Brain Imaging with focus on White Matter and Callosal Regions	MRI
CVC-ClinicDB (Dataset of Endoscopic Colonoscopy Frames for Polyp Detection)	Endoscopic Frames for Polyp Detection and Segmentation	Colonoscopy
Montgomery County Lung CT (Chest CT Segmentation) Dataset	Lung, Heart, and Trachea Segmentation from Chest CT scans	CT
Brain MRI Segmentation Dataset	White Matter, Gray Matter, and Cerebrospinal Fluid Segmentation	MRI
Chest X-ray Masks and Labels (Pulmonary Chest X-Ray Defect Detection)	Detection and Segmentation of Pulmonary Abnormalities	X-ray
Breast UltraSound Images (BUSI) Dataset	Breast Ultrasound Images for Tumor Classification	Ultrasound
(MoNuSeg) Dataset	Nuclear Segmentation in Histopathology Images	Histopathology
PhysioNet Dataset	Cardiovascular Signals and Images for Clinical Research	Multi-modality (ECG, Ultrasound, etc.)
Retina Dataset	Fundus images for Retinal Disease Detection	Fundus Photography
ChestXray Dataset	Chest X-ray Images for Pulmonary Disease Diagnosis	X-ray
LiverSeg Dataset	Liver Segmentation from CT images	CT
Heart CT Dataset	Heart Structure Segmentation from CT scans	CT
CTBrain-MRCT Dataset	Multi-modal Brain Imaging with CT and MRI	CT, MRI
MRBrain-MRCT Dataset	Brain Structure Segmentation using MR-CT Fusion	MRI, CT
ETIS-Polyp Dataset	Endoscopic Images with Annotated Polyps	Colonoscopy
PolypSeg Dataset	Segmentation Dataset for Polyp Detection	Colonoscopy
PolypDB BLI Dataset	Blue Light Imaging (BLI) for Polyp Detection	Colonoscopy
PolypDB FICE Dataset	FICE-enhanced Colonoscopy Images	Colonoscopy
PolypDB LCI Dataset	Linked Color Imaging (LCI) for Polyp Detection	Colonoscopy
PolypDB NBI Dataset	Narrow Band Imaging (NBI) for Polyp Analysis	Colonoscopy
PolypDB WLI Dataset	White Light Imaging (WLI) for Polyp Detection	Colonoscopy
BUSI Synthetic Dataset	Synthetic Breast Ultrasound Images	Ultrasound
REFUGE Dataset	Retinal Fundus Images	Fundus Photography
Abdominal OrganSeg Dataset	Abdominal Organ Segmentation from CT Images	CT
PROMISE Dataset	Prostate MRI for Segmentation and Diagnosis	MRI
ISIC Dataset	Skin Lesion Images	Dermoscopy

TABLE I: Description of the Collected Datasets

with the highest DICE score, ensuring that minor deviations across predicted masks are accounted for and the most accurate prediction is retained for further analysis.

***Mask-Matching Mechanism**:* Given a specific object (foreground) within an image, N binary predicted masks P_n are generated, each offering a unique segmentation. The DICE score between each predicted mask P_n and the ground truth (GT) mask G is calculated as:

$$DICE(P_n, G) = \frac{2|P_n \cap G|}{|P_n| + |G|}$$

The predicted mask with the highest DICE score is selected

Model	#Params (M)	#MACs (G)	Throughput (image/s)	COCO mAP
SAM-ViT-B	86	17.6	11	46.5
SAM-ViT-H	632	113.4	7	49.1
MobileSAM [20]	9.8	39	278	38.7
EdgeSAM [21]	9.6	20	449	42.1
EfficientViT-SAM-XL0 [18]	117.0	185	278	47.5
EfficientViT-SAM-XL1 [18]	203.3	322	182	47.8
EfficientSAM (proposed)	*25.3*	*247*	*183*	*44.4*

TABLE II: Runtime Efficiency Comparison. Runtime efficiency comparison was benchmarked on a single NVIDIA Tesla T4 GPU with TensorRT, fp16.

for final evaluation:

$$P_{best} = \underset{n=1,\ldots,N}{argmax}(DICE(P_n, G))$$

Where:

- P_{best} represents the optimal predicted mask for segmentation evaluation.
- P_n denotes the n^{th} predicted binary mask for the object.
- N is the total number of predicted binary masks for an object in one image.
- $\cap$ indicates the intersection operation between the predicted mask and the ground truth.
- $argmax$ returns the index of the mask with the highest DICE score.

This process ensures that only the most precise segmentation is considered during performance evaluation, eliminating the potential influence of suboptimal predictions.

Metric Computation for Segmentation Masks: To compute the relevant metrics for segmentation evaluation, Algorithm 1 (seen below) was implemented. This algorithm outlines the process of converting masks to binary, computing key metrics such as DICE Score (DSC), Jaccard Index (JAC)/Intersection over Union (IoU), and Hausdorff Distance (HD), and ensuring shape compatibility between predicted and ground truth masks.

Inference and Metric Evaluation: Additionally, the structure of the inference process can be seen in Algorithm 2. This algorithm details the steps taken during inference, including moving the model to the appropriate device, processing batches of data, and computing metrics for each prediction.

By employing these algorithms, we ensure a structured and efficient approach to segmenting and evaluating the performance of EfficientSAM.

Algorithm 1 Metric Computation for Segmentation Masks

Require: Predicted masks $P \in \mathbb{R}^{N\times C\times H\times W}$ which represents a batch of N images, where each image has C channels, and each channel has a height of H pixels, width of W pixels, ground truth masks $G \in \{0,1\}^{N\times 1\times H\times W}$, threshold θ

Convert predicted masks to binary: $P \leftarrow (P > \theta).float()$
Convert ground truth masks to binary: $G \leftarrow (G > \theta).float()$
if $C > 1$ **then**
  Flatten predicted masks: $P \leftarrow P[:, 0, :, :]$
end if
if shape of $P \neq G$ **then**
  Raise Error: Shape mismatch between predicted and ground truth masks.
end if
Flatten masks for metrics:
$P_{flat} \leftarrow P.view(P.size(0), -1)$
$G_{flat} \leftarrow G.view(G.size(0), -1)$
Compute Dice score:
$I \leftarrow \sum P_{flat} \cdot G_{flat}$ {Intersection}
$D \leftarrow \frac{2 \cdot I}{\sum P_{flat} + \sum G_{flat}}$
Compute Jaccard Index (IoU):
$U \leftarrow \sum P_{flat} + \sum G_{flat} - I$ {Union}
$J \leftarrow \frac{I}{U}$
Initialize Hausdorff Distance list: $HD \leftarrow []$
for $i = 1 \ldots N$ **do**
  $P_{coords} \leftarrow$ coordinates of non-zero pixels in $P[i]$
  $G_{coords} \leftarrow$ coordinates of non-zero pixels in $G[i]$
  if both masks have points **then**
    $hd \leftarrow \max(directed\_hausdorff(P_{coords}, G_{coords}), directed\_hausdorff(G_{coords}, P_{coords})$
  else
    $hd \leftarrow \infty$ {Assign large distance if one mask is empty}
  end if
  $HD.append(hd)$
end for
Return metrics:
return $\{D, J, \text{mean}(HD)\}$

IV. Experimental Results and Discussion

A. Implementation Details

Software Environment: The implementation of EfficientSAM utilized the following software stack: *Python* (version 3.10.12) for the programming language, *PyTorch* (version 2.4.1+cu121) and *TorchVision* (version

Algorithm 2 Inference and Metric Evaluation

Require: Model M, DataLoader D, Device d
Move model to device: $M.to(d)$
Initialize metrics list $metrics\_list \leftarrow []$
$M.eval()$ {Set model to evaluation mode}
Disable gradient calculation:
for batch in D **do**
  $X \leftarrow$ batch["pixel_values"].$to(d)$
  $G \leftarrow$ batch["ground_truth_mask"].$to(d)$
  $Y \leftarrow M(X)$ {Predict masks}
  $metrics \leftarrow compute\_metrics(Y, G, \theta = 0.4)$
  $metrics\_list.append(metrics)$
end for
Calculate average metrics:
$average\_metrics \leftarrow$ average of $metrics\_list$
return average_metrics

0.19.1+cu121) for the deep learning framework, and additional libraries adopted were *torch.cuda.amp* for mixed-precision training, *patchify* for data preprocessing, and *Matplotlib* for visualization.

***Dataset Collection and Image Preprocessing**:* The preparation involved image resizing, patch extraction, normalization, and filtering of non-informative data. A custom function standardized the dataset by loading and resizing images and masks to a uniform dimension, ensuring consistency for model integration. Images were converted to grayscale and processed for patch extraction, which improves computational efficiency by breaking down high-resolution images into manageable segments. The patch size was chosen based on model requirements and image characteristics.

Normalization was performed on mask patches to standardize pixel values, enhancing model training by reducing biases from varying intensity values. Patches with empty masks (maximum pixel value of zero) were filtered out to ensure only informative data was included, improving training robustness.

The valid image and mask patches were then organized into a structured dictionary format for efficient data management and integration into the training pipeline. This structured dataset was transformed into a compatible dataset object for the deep learning framework and serialized into a pickle file for easy access in future experiments, ensuring data integrity and quick loading times.

***Training Preparation for EfficientSAM**:* The preparation and customization of the dataset aimed to facilitate the training of EfficientSAM for medical image analysis. This involved using custom data loaders: where we implemented efficient data loading mechanisms to handle high-resolution medical images without significant bottlenecks, as well as a balanced class distribution: where we ensured a balanced representation of various classes within the dataset to mitigate class imbalance issues during training.

***Mixed-Precision Training for Fine-Tuning EfficientSAM**:* In fine-tuning EfficientSAM for medical image analysis, mixed-precision training was employed due to its substantial advantages over traditional training methods. By utilizing both 16-bit and 32-bit floating-point numbers, mixed-precision training significantly reduced memory usage and accelerated the training speed while preserving EfficientSAM's accuracy. This was especially beneficial for processing large volumes of high-resolution medical images, allowing for larger batch sizes that improved convergence and performance. Leveraging the Tesla T4 GPU architecture, mixed-precision training facilitated rapid iteration on model design and hyperparameter tuning, which is crucial for time-consuming results in medical image analysis.

The fine-tuning process began with the Adam optimizer and a custom *DiceCELoss* function, using a learning rate of 1×10^{-5} over 100 epochs on the GPU, without weight decay. During training, batches of data were processed to compute segmentation loss, enabling backpropagation and parameter updates, allowing the optimizer to update new gradients effectively, with mean loss recorded for progress tracking.

In the mixed-precision phase, gradient scaling was employed using a *GradScaler* to optimize gradients during training. Checkpoints were created for recovery, and gradients were reset before backpropagation, using scaled loss for gradient calculations. The model's state dictionary was saved post-training for evaluation and inference.

***Testing Strategy and Design**:* After preparing the test dataset, the next step involved loading the pre-trained model using the *transformers* library. The model checkpoint was accessed from a specified path, and device mapping was established to leverage GPU resources for optimized evaluation. The model's state, including weights and training scaler, was restored to maintain parameter integrity for performance assessment.

The evaluation process calculated key performance metrics, such as the Dice score, Jaccard Index (JAC/Intersection over Union - IoU), and Hausdorff Distance (HD). These metrics were derived from binary thresholds applied to predicted and ground truth masks, accurately reflecting segmentation performance. The test dataset was organized into a DataLoader for efficient batching, and the model was set to evaluation mode to optimize memory usage. The evaluation loop processed batches, conducted inference, and calculated metrics. After this process, the resulting average metrics across all batches provided valuable insights into the model's performance in medical image analysis tasks.

B. Evaluation Metrics

In evaluating the performance of segmentation models, several key metrics are employed to quantify the accuracy and effectiveness of the predictions. These metrics provide insight into how well EfficientSAM aligns with the ground truth and help identify areas for improvement.

*1) **DICE Similarity Coefficient** (DICE, %)[27], [28]:* The DICE coefficient measures the overlap between the predicted segmentation and the ground truth and is defined mathemati-

cally as:

$$Dice(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|}$$

where:

- A denotes the Segmented Region.
- B denotes the Ground Truth.
- $|A \cap B|$ denotes the magnitude of the intersection between A and B.
- $|A|$ denotes the magnitude of the segmented region.
- $|B|$ denotes the magnitude of the ground truth.

The DICE coefficient ranges from 0 (no overlap) to 1 (perfect overlap), making it a useful metric for assessing model performance, particularly in scenarios with class imbalance.

*2) **Jaccard Similarity Coefficient** (JAC, %)[29], [28]:* The Jaccard coefficient, also known as the Intersection over Union (IoU), provides a measure of the overlap between the predicted and actual segmentation. It is defined as:

$$JAC = \frac{|A \cap B|}{|A| \cup |B|}$$

where:

- A denotes the Segmented Region.
- B denotes the Ground Truth.
- $|A \cap B|$ denotes the magnitude of the intersection between A and B.
- $|A \cup B|$ denotes the magnitude of the union of A and B.

The Jaccard coefficient ranges from 0 to 1, similar to the DICE coefficient, but it may be more sensitive to differences in smaller regions.

*3) **Hausdorff Distance** (HD, pixel)[30], [28]:* The Hausdorff distance measures the maximum distance between the predicted segmentation and the ground truth, and is defined as:

$$HD(A, B) = max\{sup_{a \in A}(inf_{b \in B} d(a, b)), \\ sup_{b \in B}(inf_{a \in A} d(a, b))\}$$

where:

- A denotes the Segmented Region.
- B denotes the Ground Truth.
- d(a,b) denotes the distance metric between points a and b.
- sup denotes the Supremum (Least Upper Bound).
- inf denotes the Infimum (Greatest Lower Bound).

The Hausdorff distance is particularly useful for understanding the worst-case error between two sets of points, making it a valuable metric for segmentation tasks, especially in complex medical images.

C. Comprehensive Analysis of SAM-based Models Across Medical Modalities

Evaluation of the segmentation performance of SAM-based models using various ViT architectures across various medical imaging modalities has been provided below. The metrics used include the Dice Similarity Coefficient (DSC) [III] and Jaccard Similarity Coefficient (JAC) [IV]. These results offer a comprehensive comparison to provide insights into both the strengths and limitations of each model, with an emphasis on EfficientSAM for efficiency in medical image segmentation.

***DICE Similarity Coefficient** (DSC):* Table III shows the DSC scores across different modalities, highlighting the performance of each model.

Based on Table III, we see that EfficientSAM consistently achieves the highest DSC scores across most modalities, reflecting its versatility in handling diverse datasets. EfficientSAM performs moderately well across multiple modalities, achieving better results than SAM ViT-H but not outperforming SAM ViT-B.

EfficientSAM achieves higher DSC scores due to its optimized architecture, mixed precision training, domain-specific fine-tuning, and SAM's robust embedding framework. These factors allow it to segment medical images more accurately than traditional models, which are either too generalized or less efficient at handling complex medical data as seen above.

***Jaccard Similarity Coefficient** (JAC):* Table IV presents the JAC scores for the SAM-based models.

Using the scores from the table above, we see that EfficientSAM continues to lead with the highest JAC scores and shows promise in modalities like Fundus imaging, although further fine-tuning is required to improve performance across other datasets.

Figure 3 displays how the SAM model performs in medical images. We can see that in some of the modalities, EfficientSAM does not always produce the best results.

From the performance comparison, SAM-based methods, particularly SAM ViT-H, demonstrate superior segmentation capabilities across various modalities when in comparison with SAM ViT-B. However, EfficientSAM shows potential for improvement with additional fine-tuning. The diverse performance across different metrics highlights the trade-offs between accuracy, efficiency, and generalization.

D. Analysis of Factors Correlating to Segmentation Results

One of the major differences caused in the performance of EfficientSAM compared to other SAM variants is due to the key reason of using the *DiceCE* Loss Function, which combines the *DICE* Loss with the *Cross-Entropy* Loss.

The Dice loss measures the overlap between the predicted mask $\hat{y}$ and the ground truth mask y and focuses on how well the predicted regions align with the correct ones, especially in the presence of class imbalance. The DICE Loss Function is defined by:

$$\text{Dice Loss} = 1 - \frac{2 \sum_{i=1}^{N} \left(\hat{y}_i^2 + y_i^2\right) + \epsilon}{2 \sum_i \hat{y}_i y_i + \epsilon}$$

Where:

- $\hat{y}_i$ represents the predicted probability for the i-th pixel.

Method	Endoscopy	Colonoscopy	Ultrasound	Cardiac	Fundus	X-ray	CT	MRI	Dermoscopy	Avg.
SAM-based methods										
SAM ViT-B [31]	0.0018	0.0128	0.0074	0.0105	0.33	0.0118	0.0066	0.0067	0.0212	0.045
SAM ViT-H [31]	0.0021	0.0193	0.0099	0.0109	0.267	0.0484	0.008	0.008	0.0247	0.049
EfficientSAM (Proposed)	**0.0381**	**0.2129**	**0.10**	**0.0174**	**0.785**	**0.162**	**0.0558**	**0.00851**	**0.183**	**0.174**

TABLE III: Dice Similarity Coefficient (DSC) for Selected Segment Anything Model Methods

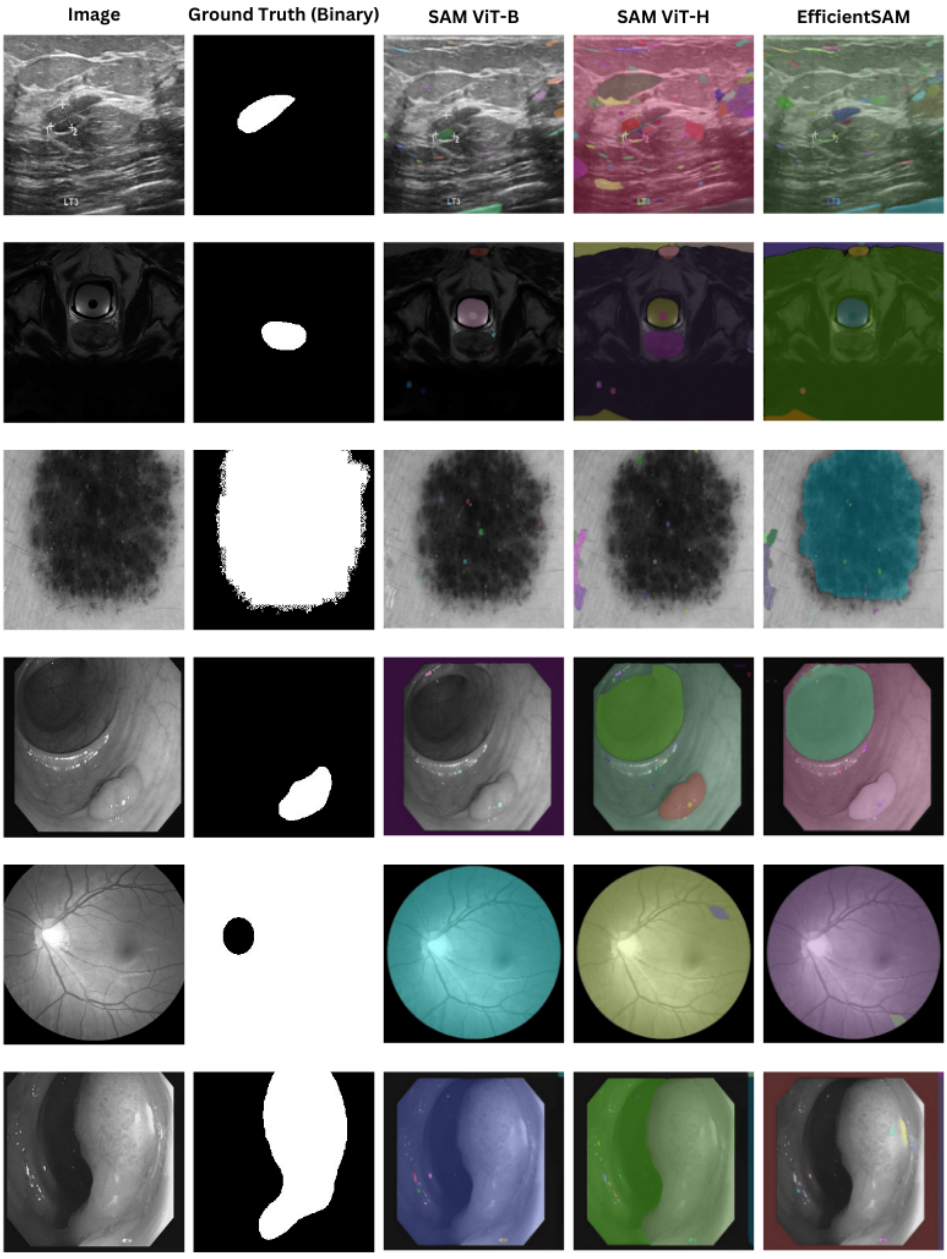

Fig. 3: Qualitative comparison of segmentation performance among SAM ViT-B, SAM ViT-H, and EfficientSAM on some of the image modalities from the test dataset.

- y_i represents the ground truth value (binary: 0 or 1) for the i-th pixel.
- N represents the total number of samples (or pixels).
- ϵ represents a small constant to avoid division by zero.
- If *squared_pred=True*, the predictions are squared, hence enhancing the influence of larger prediction errors.

The Cross-Entropy Loss measures how well the predicted probabilities match the actual class labels/ground truth. It

Method	Endoscopy	Colonoscopy	Ultrasound	Cardiac	Fundus	X-ray	CT	MRI	Dermoscopy	Avg.
SAM-based methods										
SAM ViT-B [31]	0.0009	0.0076	0.0044	0.0057	0.273	0.0072	0.004	0.0028	0.0159	0.036
SAM ViT-H [31]	0.0011	0.0122	0.0059	0.0058	0.222	0.0318	0.0047	0.0032	0.0166	0.038
EfficientSAM (Proposed)	**0.0197**	**0.13**	**0.0598**	**0.0094**	**0.665**	**0.108**	**0.0332**	**0.0045**	**0.128**	**0.129**

TABLE IV: Jaccard Similarity Coefficient (JAC) for Selected Segment Anything Model Methods

encourages the predicted probabilities to match the correct labels by minimizing the negative log-likelihood. The Cross-Entropy Loss Function is defined by:

$$\text{CE Loss} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right)$$

Where:

- $\hat{y}_i$ represents the predicted probability from the model for the i-th sample.
- y_i represents the ground truth label (binary: 0 or 1) for the i-th sample.
- N represents the total number of samples (or pixels).
- ϵ represents a small constant to avoid division by zero.
- If *squared_pred=True*, the predictions are squared, hence enhancing the influence of larger prediction errors.

The *DiceCE* Loss Function combines these two loss functions in the following way:

$$\text{DiceCELoss} = \alpha \times DICELoss + \beta \times CELoss$$

By default, *MONAI* (Medical Open Network for AI) assigns equal weights and this was the implemented weighting applied to EfficientSAM, i.e., $\alpha = 0.5$ and $\beta = 0.5$, but these can be adjusted.

E. Comparison between EfficientSAM and Interactive Methods

EfficientSAM's performance on the evaluation metrics used can be highlighted in the following way: **DSC** measures the overlap between predicted and ground truth masks, to achieve a good DSC score, minimize false positives/negatives, precise boundary detection. **JAC** measures the ratio of intersection to union between predicted and ground truth, to achieve a good JAC score, we need to implement better feature extraction and segmentation consistency. **HD** measures the largest boundary mismatch, and to achieve a good HD score, we need to implement fine boundary alignment, boundary-aware loss, or post-processing

Several factors contribute to obtaining these high-quality metrics. Accurate detection of relevant features (e.g., boundaries, shapes, textures) from input images is essential for good segmentation. Models with strong encoders and attention mechanisms (like Vision Transformers) tend to extract better semantic and spatial information, resulting in more accurate predictions.

DSC and JAC are sensitive to the overlap between the predicted mask and the ground truth. Sharp boundary prediction is critical, especially for medical images with fine structures to achieve high metrics. Using loss functions such as Dice Loss, Boundary Loss, or Focal Loss helps EfficientSAM learn precise boundaries. Mixed precision training as well as transfer learning from pre-trained SAM ViT models help EfficientSAM to converge faster and generalize better, hence improving the DSC and JAC scores from the model.

Other factors that would lead to well-performing segmentation models is having balanced data across classes. This ensures that the model does not overfit to one particular region or shape, improving performance across all segments. Proper generalization ensures EfficientSAM harnessing the existing ViT architecture performs well even on images from new modalities or datasets, boosting EfficientSAM's DSC and IoU scores, thus leading to better results.

The use of batch processing and the data pre-proecessing techniques used for the combined dataset removing noise and artifacts from modalities like ultrasound or MRI, this help boosted a better overlap with ground truth, leading to improved DSC and JAC scores.

V. Conclusion and Future Work

Conclusion

EfficientSAM, while designed to be a lightweight and efficient segmentation model, exhibits certain weaknesses, particularly in its initial ability to accurately perceive and segment complex medical objects. These limitations often stem from the model's training, in which the combined dataset may not adequately represent all or specific features and variations present in medical imaging. As a result, the model performs decently well, showing improvements from SAM ViT-B and SAM ViT-H, but diverges when it came to measuring distances between segmented boundary and the true boundary, leading to an overall decent increase in segmentation performance.

The impact of fine-tuning, mixed-precision and batch processing is evidenced by a notable improvement in segmentation performance metrics. For example, after fine-tuning, the Dice Similarity Coefficient (DSC) and Jaccard Similarity Coefficient (JAC) scores are often significantly higher, indicating that the model now achieves better overlap between predicted and ground truth masks. Additionally, the Hausdorff Distance (HD), which measures the maximum boundary discrepancy, tends to decrease, suggesting that the model's predictions align more closely with the actual object boundaries.

By exposing the model to a diverse set of annotated medical images, it can learn to identify intricate details, variations in appearance, and the context of objects more effectively. This targeted training allows EfficientSAM to better generalize

across different medical scenarios and improve its performance on tasks that were previously challenging.

Future Work

Incorporating data augmentation techniques like rotation, flipping, and scaling during training can enhance the robustness of EfficientSAM. These methods create a diverse training dataset, allowing the model to learn invariant features essential for accurate segmentation. As a result, EfficientSAM effectively handle unseen data, reduce overfitting, and improve generalization even further.

EfficientSAM employs architectures that balance depth and efficiency, similar to UNet++ [32], EfficientNet [33], and other SAM-based models. This design captures both local patterns, crucial for fine details in medical images, and global patterns for understanding broader anatomical structures, leading to improved segmentation accuracy.

To mitigate overfitting, regularization techniques such as dropout and batch normalization can be utilized. Early stopping methods can also help by halting training when validation performance declines. These strategies work together to enhance generalization and improve segmentation metrics across various datasets.

Additionally, post-processing techniques like thresholding, Conditional Random Fields (CRFs), and morphological operations refine segmentation outputs. Thresholding adjusts predicted probabilities for more accurate binary masks, CRFs enhance spatial coherence, and morphological operations clean up outputs. Together, these refinements ensure clinically relevant segmentation results that would reflect the effectiveness of the training strategies implemented.

REFERENCES

[1] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[2] Y. Huang, X. Yang, L. Liu, H. Zhou, A. Chang, X. Zhou, R. Chen, J. Yu, J. Chen, C. Chen, *et al.*, "Segment anything model for medical images?" *Medical Image Analysis*, vol. 92, p. 103061, 2024.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[4] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.

[5] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.

[6] L. Tang, H. Xiao, and B. Li, "Can sam segment anything? when sam meets camouflaged object detection," *arXiv preprint arXiv:2304.04709*, 2023.

[7] W. Ji, J. Li, Q. Bi, T. Liu, W. Li, and L. Cheng, "Segment anything is not always perfect: An investigation of sam on different real-world applications," 2024.

[8] G.-P. Ji, D.-P. Fan, P. Xu, M.-M. Cheng, B. Zhou, and L. Van Gool, "Sam struggles in concealed scenes–empirical study on," *segment anything*, 2023.

[9] S. Mohapatra, A. Gosai, and G. Schlaug, "Sam vs bet: A comparative study for brain extraction and segmentation of magnetic resonance images using deep learning," *arXiv preprint arXiv:2304.04738*, 2023.

[10] R. Deng, C. Cui, Q. Liu, T. Yao, L. W. Remedios, S. Bao, B. A. Landman, L. E. Wheless, L. A. Coburn, K. T. Wilson, *et al.*, "Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging," *arXiv preprint arXiv:2304.04155*, 2023.

[11] T. Zhou, Y. Zhang, Y. Zhou, Y. Wu, and C. Gong, "Can sam segment polyps?" *arXiv preprint arXiv:2304.07583*, 2023.

[12] Y. Liu, J. Zhang, Z. She, A. Kheradmand, and M. Armand, "Samm (segment any medical model): A 3d slicer integration to sam," *arXiv preprint arXiv:2304.05622*, 2023.

[13] S. Pieper, M. Halle, and R. Kikinis, "3d slicer," in *2004 2nd IEEE international symposium on biomedical imaging: nano to macro (IEEE Cat No. 04EX821)*. IEEE, 2004, pp. 632–635.

[14] S. He, R. Bao, J. Li, J. Stout, A. Bjornerud, P. E. Grant, and Y. Ou, "Computer-vision benchmark segment-anything model (sam) in medical images: Accuracy in 12 datasets," *arXiv preprint arXiv:2304.09324*, 2023.

[15] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: an experimental study," *Medical Image Analysis*, vol. 89, p. 102918, 2023.

[16] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[17] J. Wu, W. Ji, Y. Liu, H. Fu, M. Xu, Y. Xu, and Y. Jin, "Medical sam adapter: Adapting segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.12620*, 2023.

[18] Z. Zhang, H. Cai, and S. Han, "Efficientvit-sam: Accelerated segment anything model without performance loss," *arXiv preprint arXiv:2402.05008*, 2024.

[19] C. Mattjie, L. V. De Moura, R. Ravazio, L. Kupssinskü, O. Parraga, M. M. Delucis, and R. C. Barros, "Zero-shot performance of the segment anything model (sam) in 2d medical imaging: A comprehensive evaluation and practical guidelines," in *2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2023, pp. 108–112.

[20] C. Zhang, D. Han, Y. Qiao, J. U. Kim, S.-H. Bae, S. Lee, and C. S. Hong, "Faster segment anything: Towards lightweight sam for mobile applications," *arXiv preprint arXiv:2306.14289*, 2023.

[21] C. Zhou, X. Li, C. C. Loy, and B. Dai, "Edgesam: Prompt-in-the-loop distillation for on-device deployment of sam," *arXiv preprint arXiv:2312.06660*, 2023.

[22] Y. Xiong, B. Varadarajan, L. Wu, X. Xiang, F. Xiao, C. Zhu, X. Dai, D. Wang, F. Sun, F. Iandola, *et al.*, "Efficientsam: Leveraged masked image pretraining for efficient segment anything," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 111–16 121.

[23] H. Shu, W. Li, Y. Tang, Y. Zhang, Y. Chen, H. Li, Y. Wang, and X. Chen, "Tinysam: Pushing the envelope for efficient segment anything model," *arXiv preprint arXiv:2312.13789*, 2023.

[24] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, *et al.*, "Mixed precision training," *arXiv preprint arXiv:1710.03740*, 2017.

[25] S. Nandakumar, M. Le Gallo, C. Piveteau, V. Joshi, G. Mariani, I. Boybat, G. Karunaratne, R. Khaddam-Aljameh, U. Egger, A. Petropoulos, *et al.*, "Mixed-precision deep learning based on computational memory," *Frontiers in neuroscience*, vol. 14, p. 406, 2020.

[26] M. Dörrich, M. Fan, and A. M. Kist, "Impact of mixed precision techniques on training and inference efficiency of deep neural networks," *IEEE Access*, vol. 11, pp. 57 627–57 634, 2023.

[27] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[28] P. Bilic, P. Christ, H. B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, A. Szeskin, C. Jacobs, G. E. H. Mamani, G. Chartrand, *et al.*, "The liver tumor segmentation benchmark (lits)," *Medical Image Analysis*, vol. 84, p. 102680, 2023.

[29] J. Qian, R. Li, X. Yang, Y. Huang, M. Luo, Z. Lin, W. Hong, R. Huang, H. Fan, D. Ni, *et al.*, "Hasa: hybrid architecture search with aggregation strategy for echinococcosis classification and ovary segmentation in ultrasound images," *Expert Systems with Applications*, vol. 202, p. 117242, 2022.

[30] D. Karimi and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Transactions on medical imaging*, vol. 39, no. 2, pp. 499–513, 2019.

[31] Meta AI Research, "Segment anything model checkpoints," https://github.com/facebookresearch/segment-anything?tab=readme-ov-file#model-checkpoints.

[32] A. M. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. S. Khan, "Unetr++: delving into efficient and accurate 3d medical image segmentation," *IEEE Transactions on Medical Imaging*, 2024.

[33] H. A. Shah, F. Saeed, S. Yun, J.-H. Park, A. Paul, and J.-M. Kang, "A robust approach for brain tumor detection in magnetic resonance images using finetuned efficientnet," *Ieee Access*, vol. 10, pp. 65 426–65 438, 2022.