Towards a Foundation Model for Medical Images

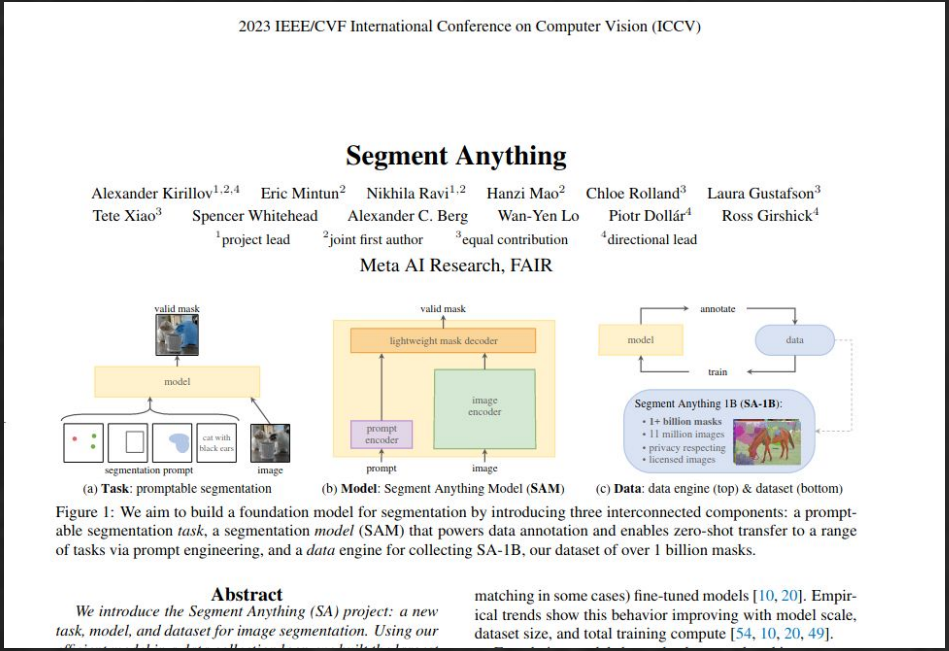
Mohamed Masoud

TReNDS Reading-Group May 2024

Background

- Segment Anything Model (SAM) Kirillov et al. [2023] is proposed, which is trained on a novel large visual dataset with 1 billion segmentation masks, termed as SA-1B.

- The goal is to build a foundation model for image segmentation that generalizes to new types of objects and images beyond what it observed during training.



2023 IEEE/CVF International Conference on Computer Vision (ICCV)

Segment Anything

Alexander Kirillov[1,2,4] Eric Mintun[2] Nikhila Ravi[1,2] Hanzi Mao[2] Chloe Rolland[3] Laura Gustafson[3]
Tete Xiao[3] Spencer Whitehead Alexander C. Berg Wan-Yen Lo Piotr Dollár[4] Ross Girshick[4]

[1]project lead [2]joint first author [3]equal contribution [4]directional lead

Meta AI Research, FAIR

(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

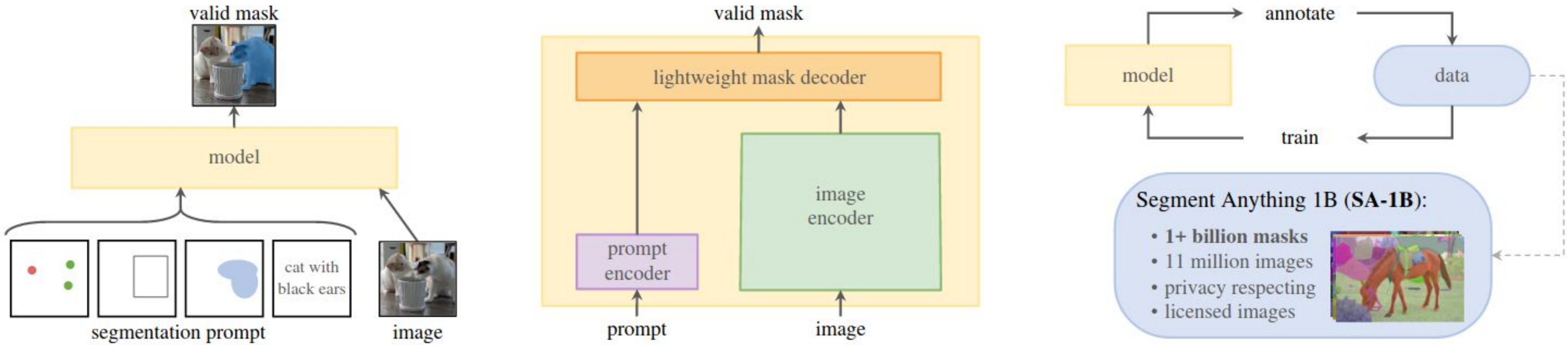
Figure 1: We aim to build a foundation model for segmentation by introducing three interconnected components: a promptable segmentation *task*, a segmentation *model* (SAM) that powers data annotation and enables zero-shot transfer to a range of tasks via prompt engineering, and a *data* engine for collecting SA-1B, our dataset of over 1 billion masks.

Abstract

We introduce the Segment Anything (SA) project: a new task, model, and dataset for image segmentation. Using our

matching in some cases) fine-tuned models [10, 20]. Empirical trends show this behavior improving with model scale, dataset size, and total training compute [54, 10, 20, 49].

SAM Components



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

Training with Simulated Prompts

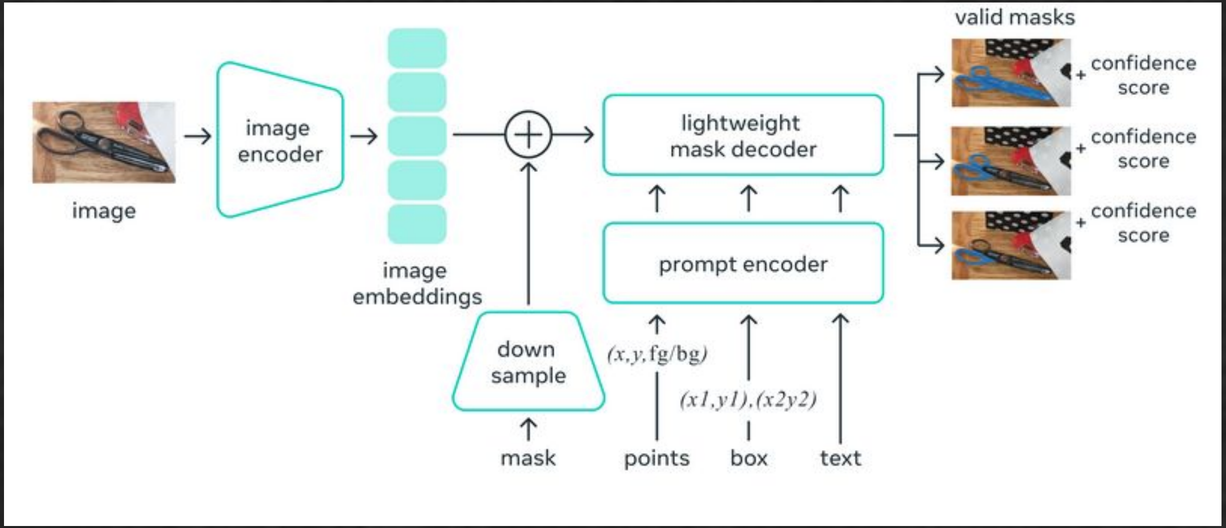
The training of the SAM model with simulated prompts is a key feature that enables it to understand and respond to various input types for segmentation tasks:

1. **Simulated Prompt Generation**: Mimic potential inputs . Prompts could be a point on an object, a bounding box surrounding an area, or even a textual description. Each prompt is associated with a specific training image.

2. **Pairing with Ground Truth**: Each simulated prompt is paired with a "ground truth" mask, which is the correct segmentation for the prompt according to human annotation. This pairing forms the basis for supervised learning, where the model learns to predict segmentation masks that match these ground truths.

3. **Training Objective**: Achieved using loss functions that measure the accuracy of the pixel-wise predictions, dice loss + cross-entropy (focal) loss.

SAM Overview

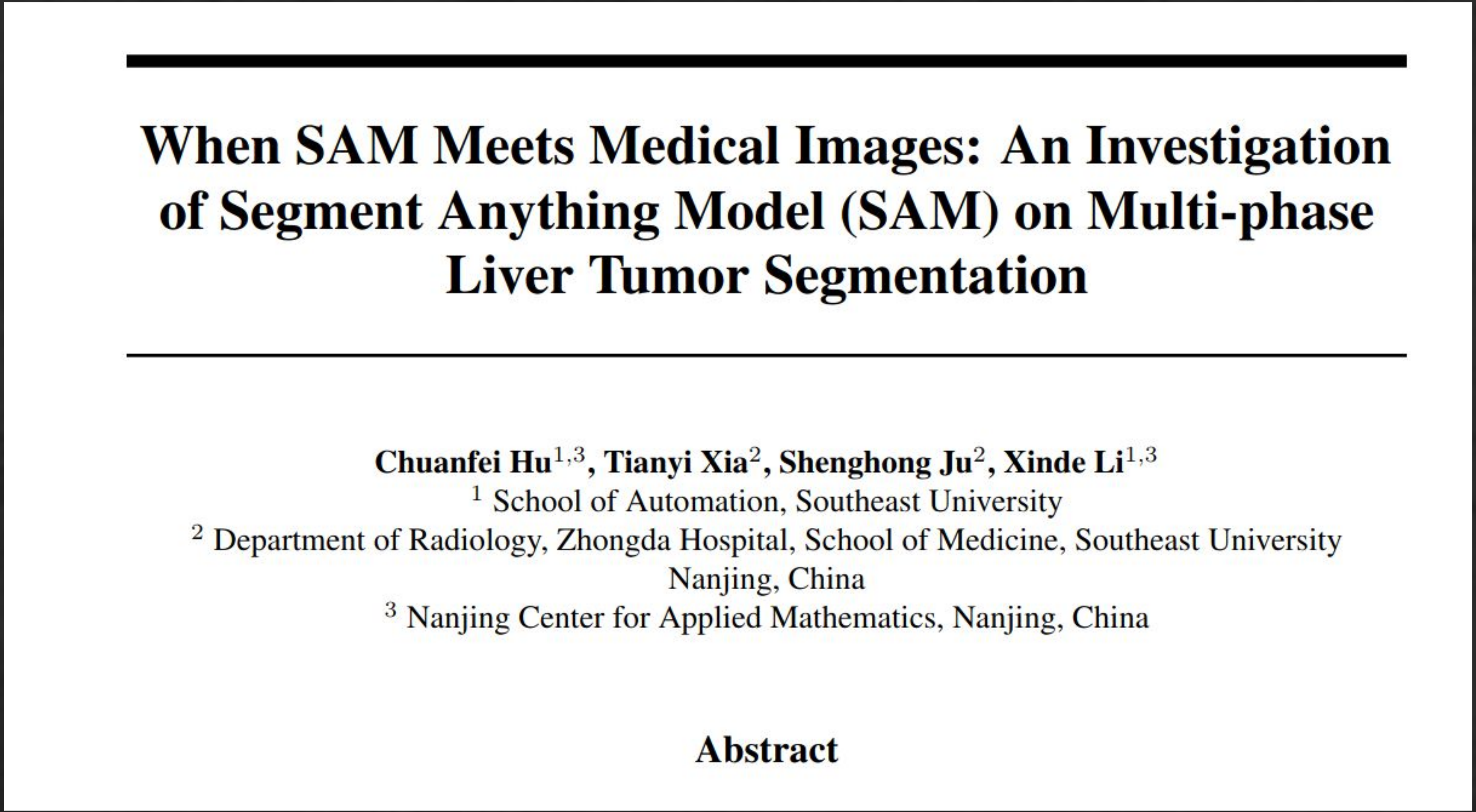


Data Engine

Assisted-manual stage: At the start of this stage, SAM was trained using common public segmentation datasets. After sufficient data annotation, SAM was retrained using only newly annotated masks.

Semi-automatic stage: Authors presented annotators with images prefilled with existing masks and asked them to annotate any additional unannotated objects. To detect confident masks, they trained a bounding box detector on all first stage masks using a generic “object” category.

Fully automatic stage: Specifically, they prompted the model with a 32⇥32 regular grid of points and for each point predicted a set of masks that may correspond to valid objects. Developed ambiguity-aware model in which if a point lies on a part or subpart, the model will return the subpart, part, and whole object. The IoU prediction module of the model is used to select confident masks. They applied fully automatic mask generation to all 11M images in the dataset, producing a total of 1.1B high-quality masks.

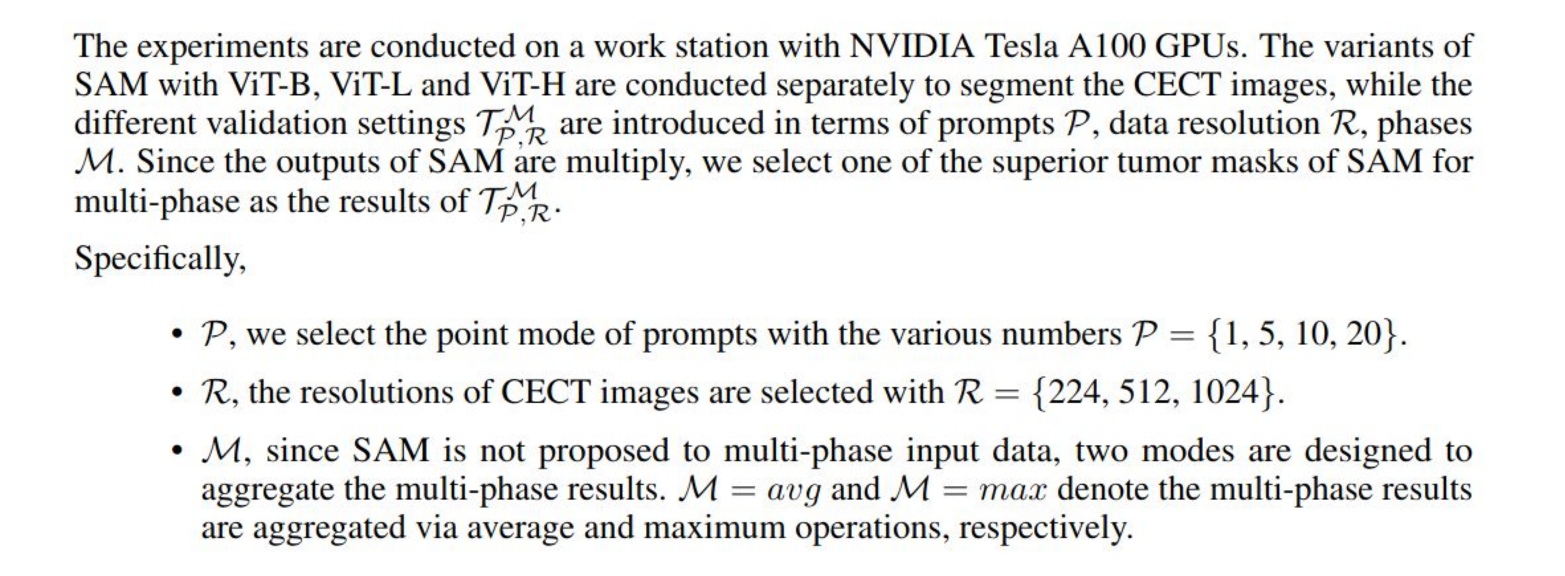


When SAM Meets Medical Images: An Investigation of Segment Anything Model (SAM) on Multi-phase Liver Tumor Segmentation

Chuanfei Hu[1,3], **Tianyi Xia**[2], **Shenghong Ju**[2], **Xinde Li**[1,3]
[1] School of Automation, Southeast University
[2] Department of Radiology, Zhongda Hospital, School of Medicine, Southeast University
Nanjing, China
[3] Nanjing Center for Applied Mathematics, Nanjing, China

Abstract

Paper Experiment

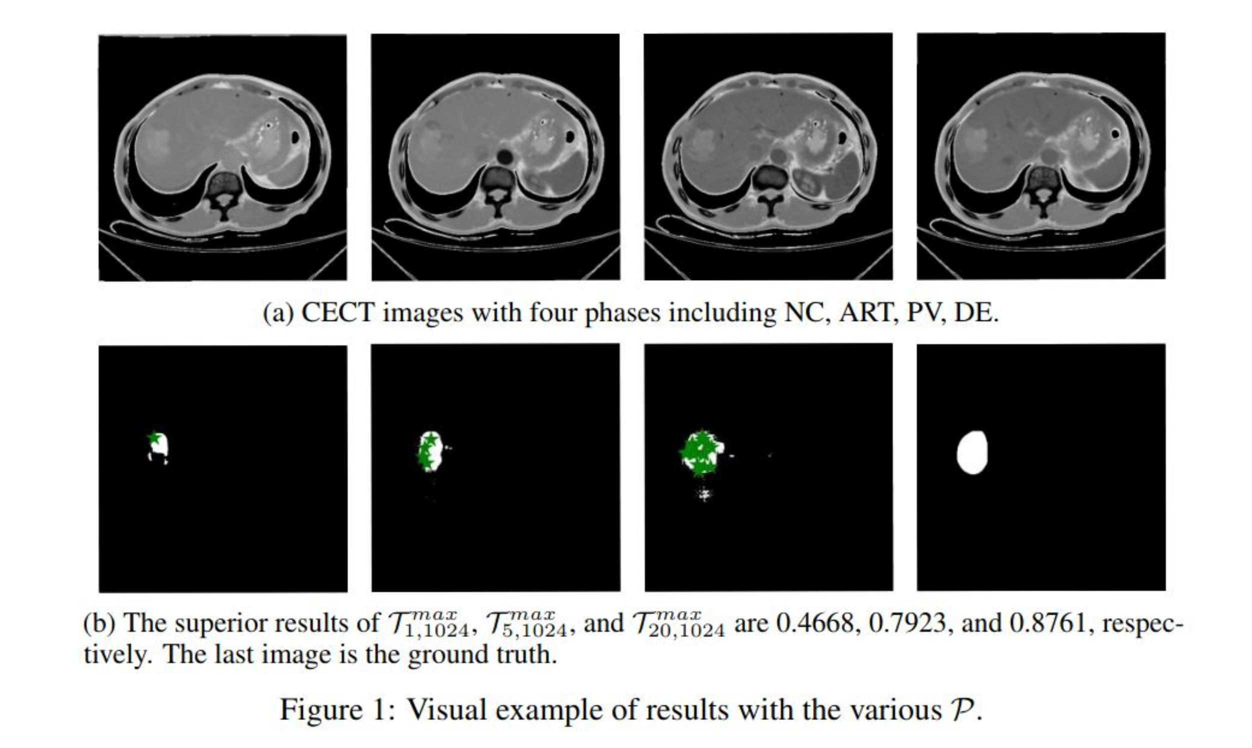


The experiments are conducted on a work station with NVIDIA Tesla A100 GPUs. The variants of SAM with ViT-B, ViT-L and ViT-H are conducted separately to segment the CECT images, while the different validation settings $\mathcal{T}^{\mathcal{M}}_{\mathcal{P},\mathcal{R}}$ are introduced in terms of prompts $\mathcal{P}$, data resolution $\mathcal{R}$, phases $\mathcal{M}$. Since the outputs of SAM are multiply, we select one of the superior tumor masks of SAM for multi-phase as the results of $\mathcal{T}^{\mathcal{M}}_{\mathcal{P},\mathcal{R}}$.

Specifically,

- $\mathcal{P}$, we select the point mode of prompts with the various numbers $\mathcal{P} = \{1, 5, 10, 20\}$.
- $\mathcal{R}$, the resolutions of CECT images are selected with $\mathcal{R} = \{224, 512, 1024\}$.
- $\mathcal{M}$, since SAM is not proposed to multi-phase input data, two modes are designed to aggregate the multi-phase results. $\mathcal{M} = avg$ and $\mathcal{M} = max$ denote the multi-phase results are aggregated via average and maximum operations, respectively.

Accuracy Vs Segmentation Points

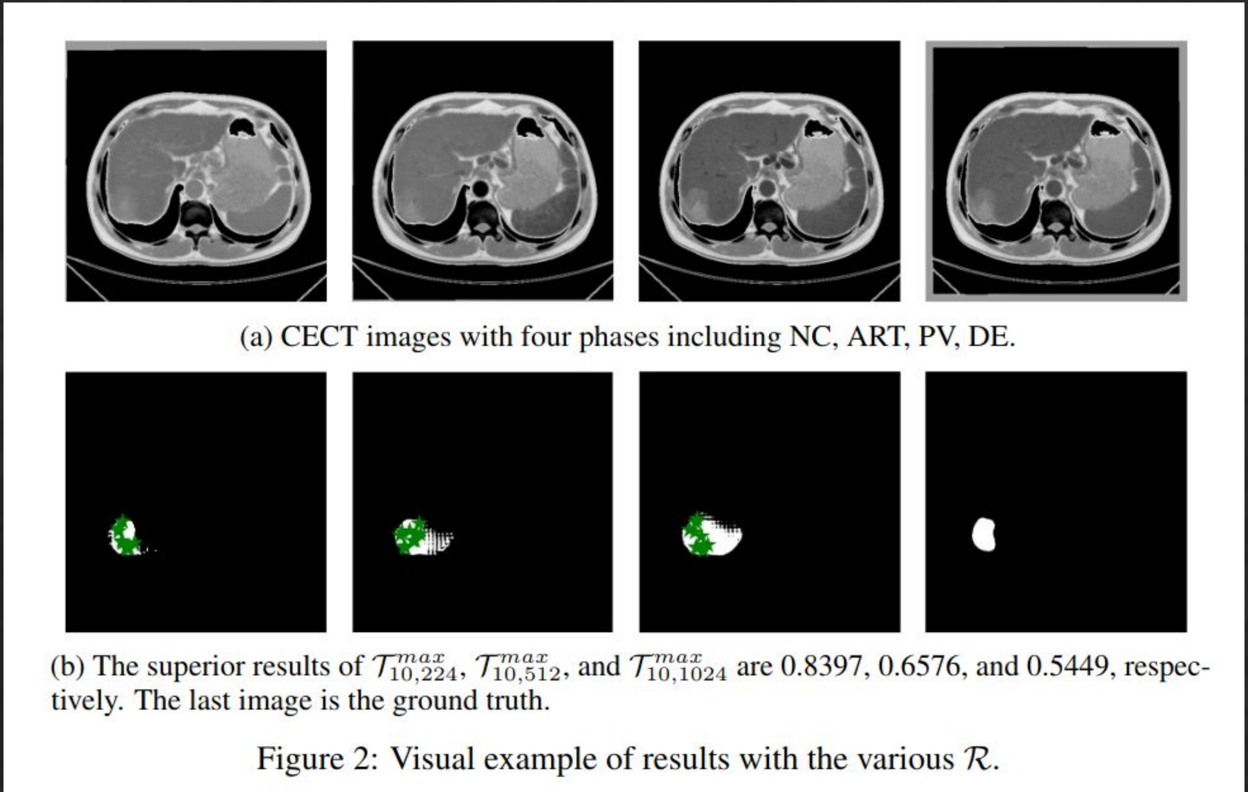


(a) CECT images with four phases including NC, ART, PV, DE.

(b) The superior results of $\mathcal{T}^{max}_{1,1024}$, $\mathcal{T}^{max}_{5,1024}$, and $\mathcal{T}^{max}_{20,1024}$ are 0.4668, 0.7923, and 0.8761, respectively. The last image is the ground truth.

Figure 1: Visual example of results with the various $\mathcal{P}$.

Accuracy Vs Resolutions

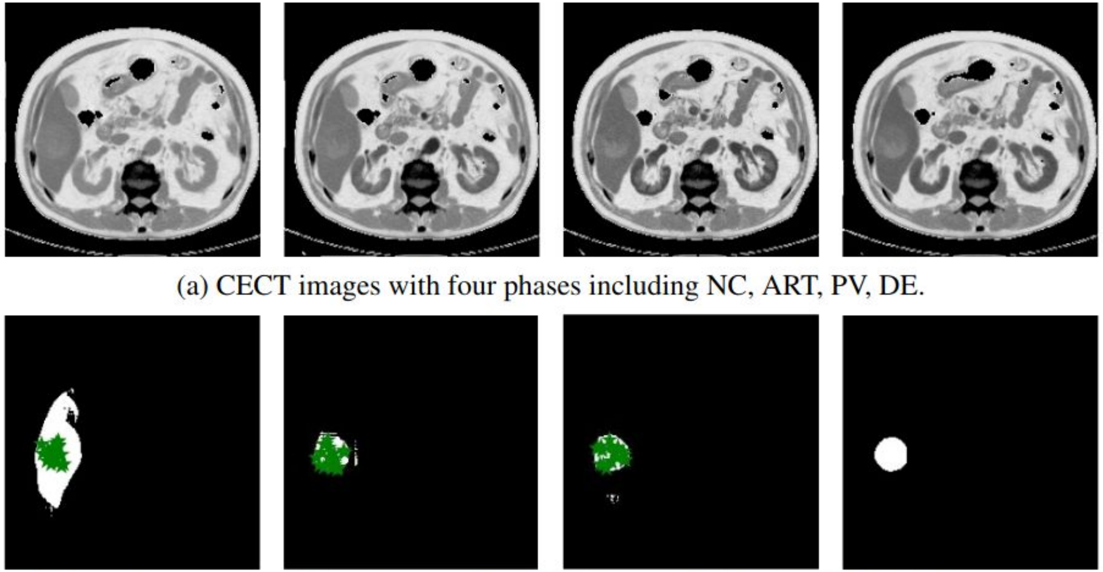


(a) CECT images with four phases including NC, ART, PV, DE.

(b) The superior results of $\mathcal{T}^{max}_{10,224}$, $\mathcal{T}^{max}_{10,512}$, and $\mathcal{T}^{max}_{10,1024}$ are 0.8397, 0.6576, and 0.5449, respectively. The last image is the ground truth.

Figure 2: Visual example of results with the various $\mathcal{R}$.

Accuracy Vs Transformers

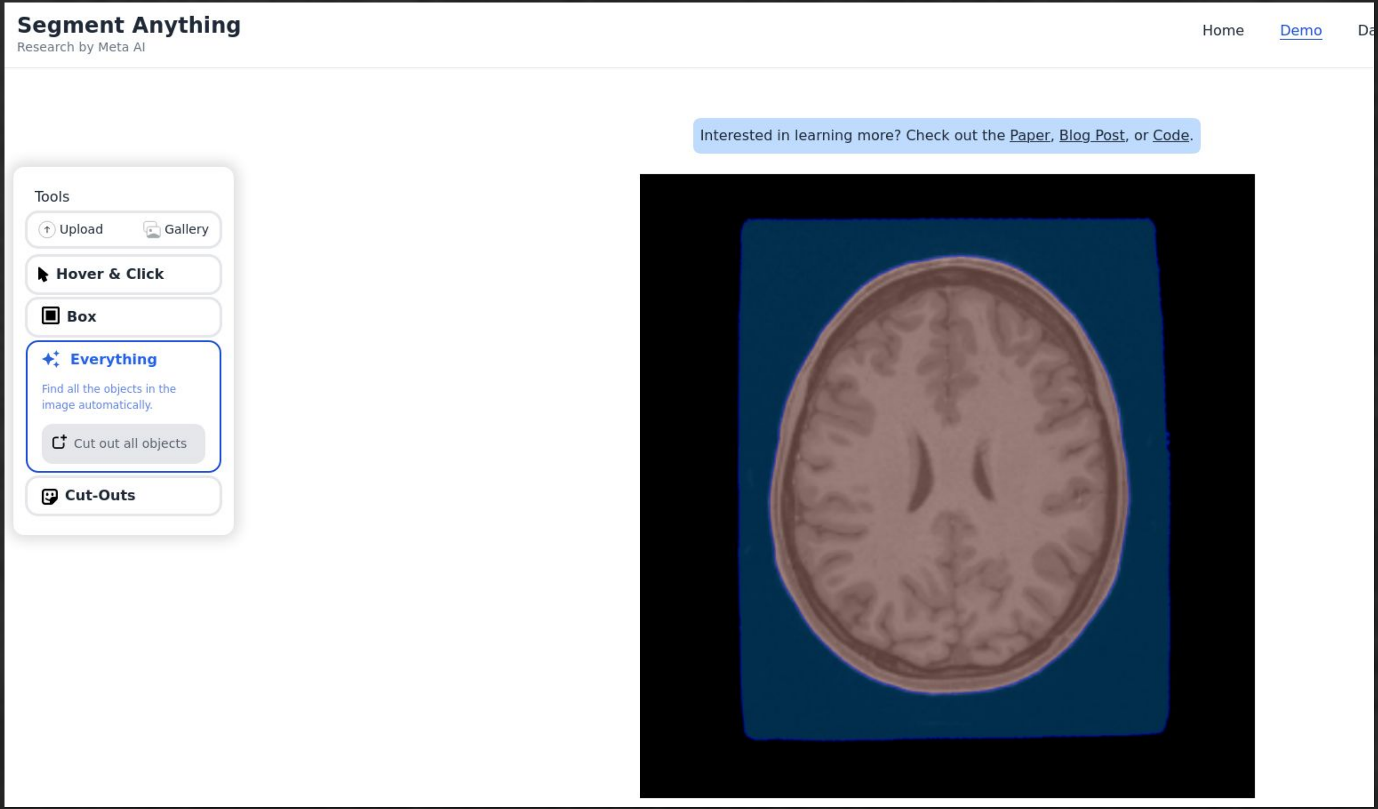


(a) CECT images with four phases including NC, ART, PV, DE.

(b) The superior results of $\mathcal{T}_{20,224}^{max}$ with ViT-b, ViT-l, and ViT-h are 0.4035, 0.8920, and 0.9246, respectively. The last image is the ground truth.

Figure 3: Visual example of results with the various $\mathcal{R}$.

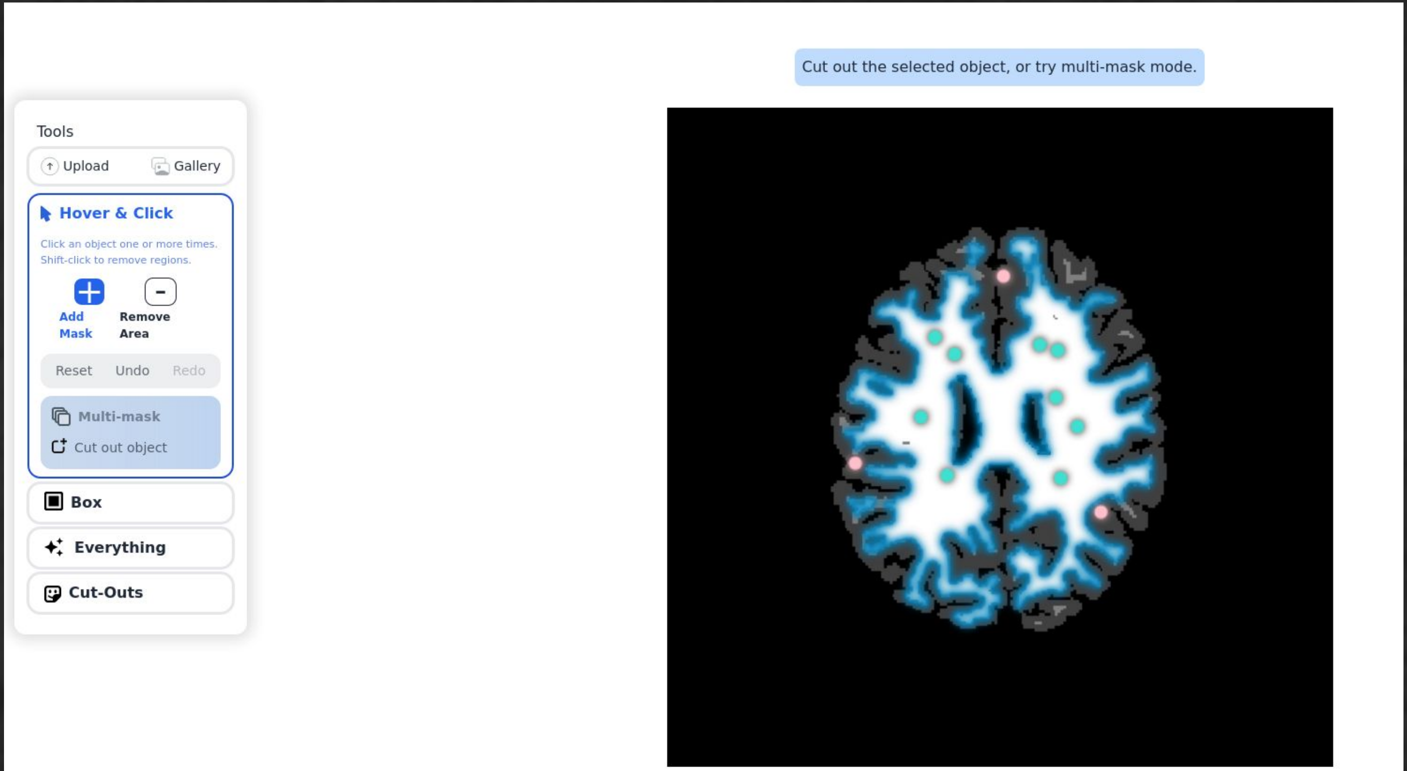
SAM Demo



Interested in learning more? Check out the Paper, Blog Post, or Code.

Find all the objects in the image automatically.

SAM Demo



Cut out the selected object, or try multi-mask mode.

Click an object one or more times. Shift-click to remove regions.

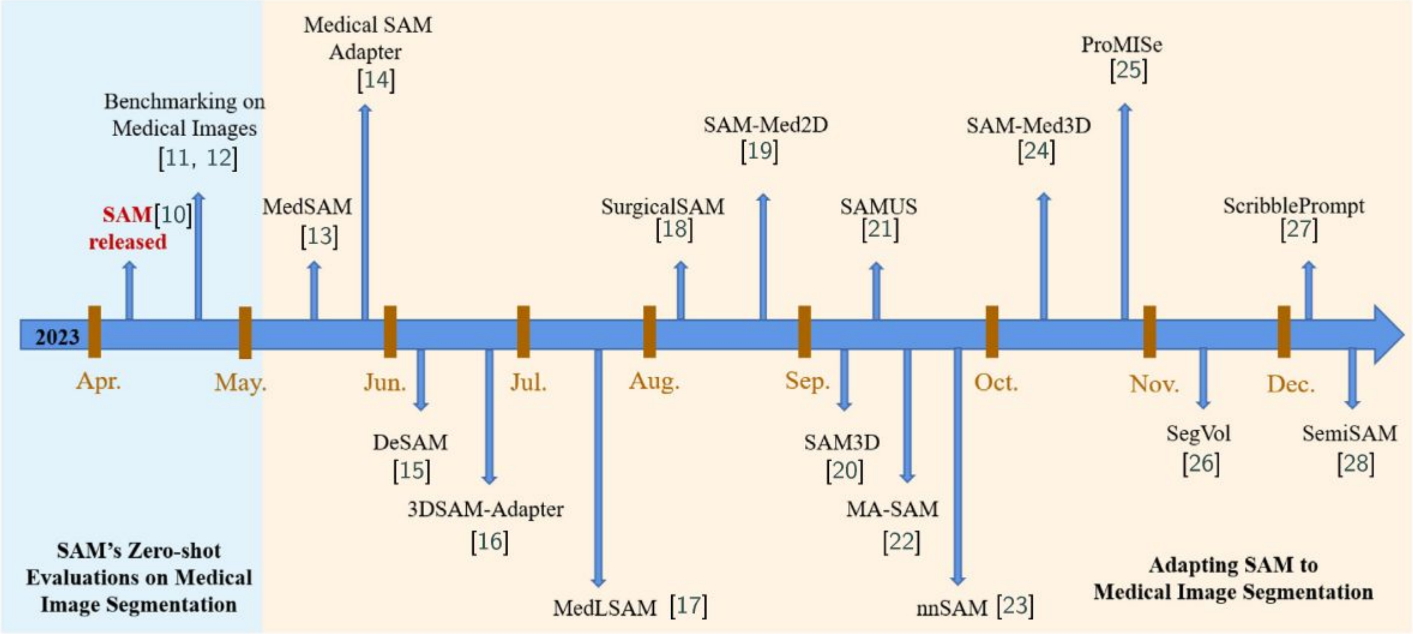
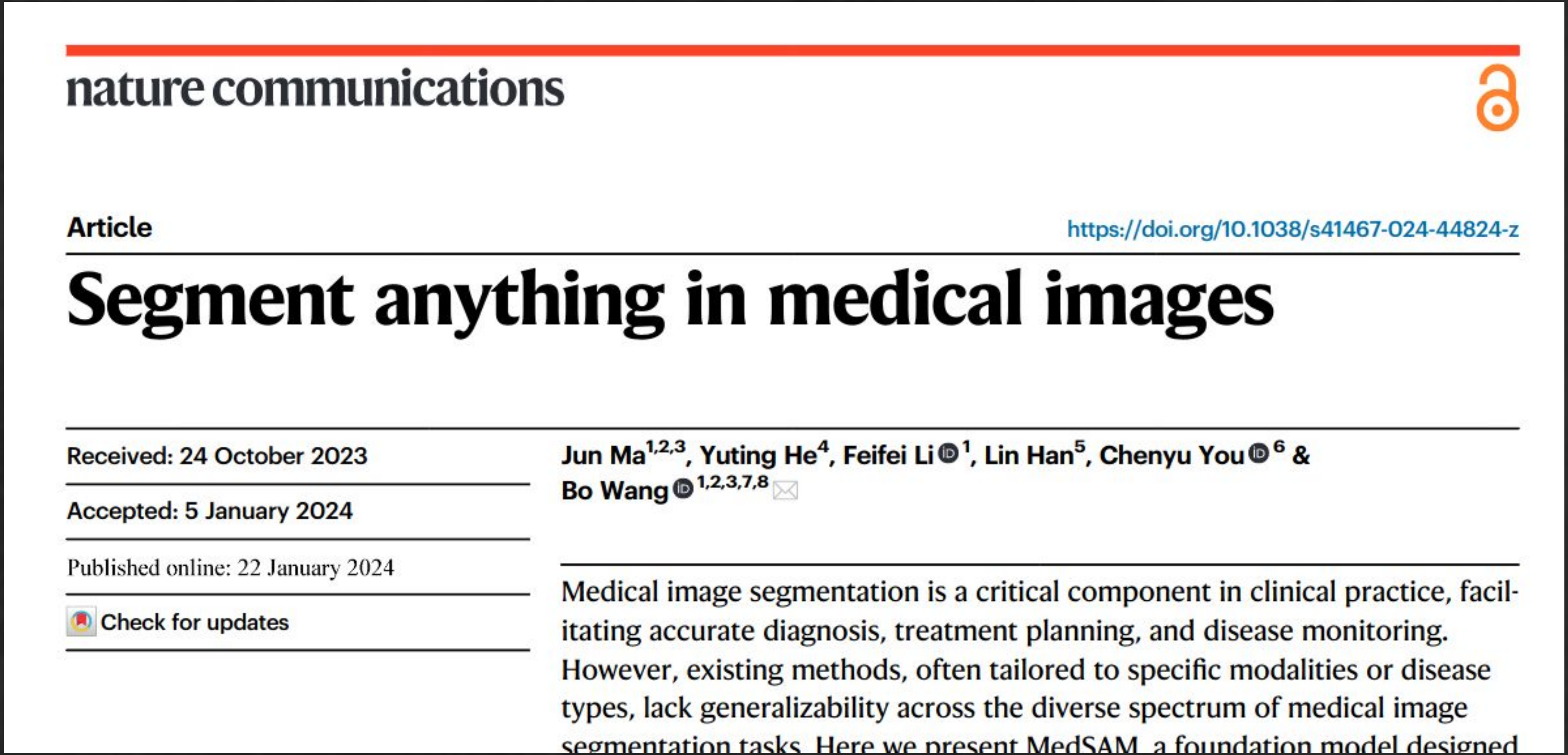


Fig. 1. A brief chronology of Segment Anything Model (SAM) [10] and its variants for medical image segmentation in 2023.



nature communications

Article

https://doi.org/10.1038/s41467-024-44824-z

Segment anything in medical images

Received: 24 October 2023

Accepted: 5 January 2024

Published online: 22 January 2024

Check for updates

Jun Ma[1,2,3], Yuting He[4], Feifei Li[1], Lin Han[5], Chenyu You[6] & Bo Wang[1,2,3,7,8]

Medical image segmentation is a critical component in clinical practice, facilitating accurate diagnosis, treatment planning, and disease monitoring. However, existing methods, often tailored to specific modalities or disease types, lack generalizability across the diverse spectrum of medical image segmentation tasks. Here we present MedSAM, a foundation model designed

Motivation

- A significant limitation of many current medical image segmentation models is their task-specific nature.
- Due to its capability of zero-shot image segmentation with some prompts, SAM is attractive, particularly for medical image analysis, where the annotations and samples are rare and laborious (C. Hu et al.)
    - However, the applicability of the segmentation foundation models (e.g., SAM[7]) to medical image segmentation remains limited due to the significant differences between natural images and medical images.
    - Also, the model (SAM) exhibited substantial limitations in segmenting typical medical targets with weak boundaries or low contrast

MedSAM idea is fine tuning SAM

1. **Fine-Tuning SAM on Medical Images**: This involves training SAM further on a dataset of medical images.

   ○ **Full Fine-Tuning**: This method involves adjusting all the model's parameters during the additional training phase on medical images. It's a comprehensive approach where the entire model (all layers and weights) is updated to better adapt to the medical imaging domain. This could lead to significant improvements in performance but requires substantial computational resources and can risk overfitting if not managed properly.

   ○ **Parameter-Efficient Fine-Tuning**: In contrast to full fine-tuning, this approach adjusts only a small subset of the model's parameters.

Challenges in Medical Images

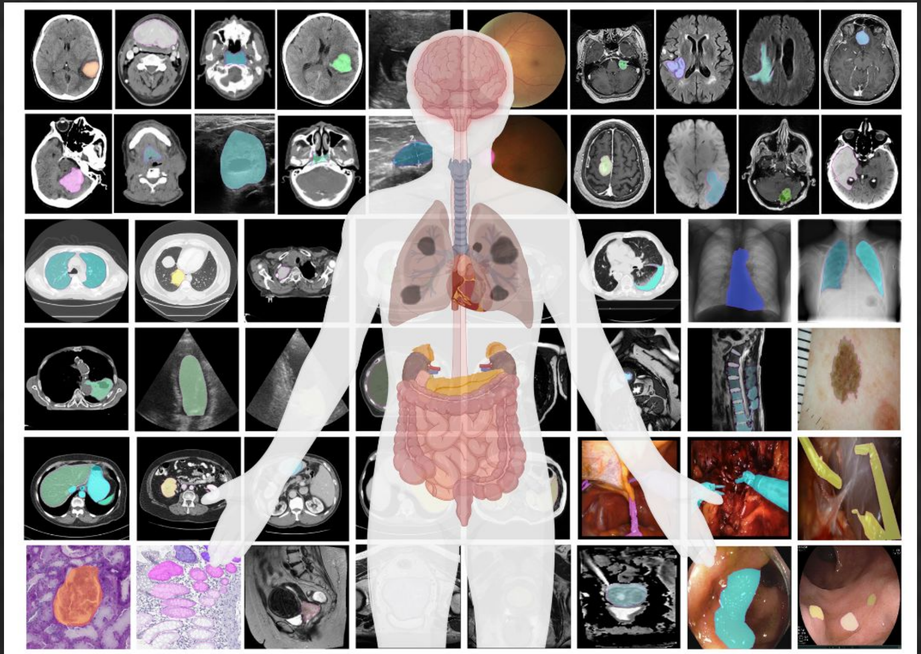
- The variability inherent in segmentation tasks.

- The variability in imaging modalities

MedSAM

In response to the limitations of general models like SAM, MedSAM has been developed. This refined model improves segmentation performance on medical images by fine-tuning on a large dataset and demonstrates superior versatility and accuracy.

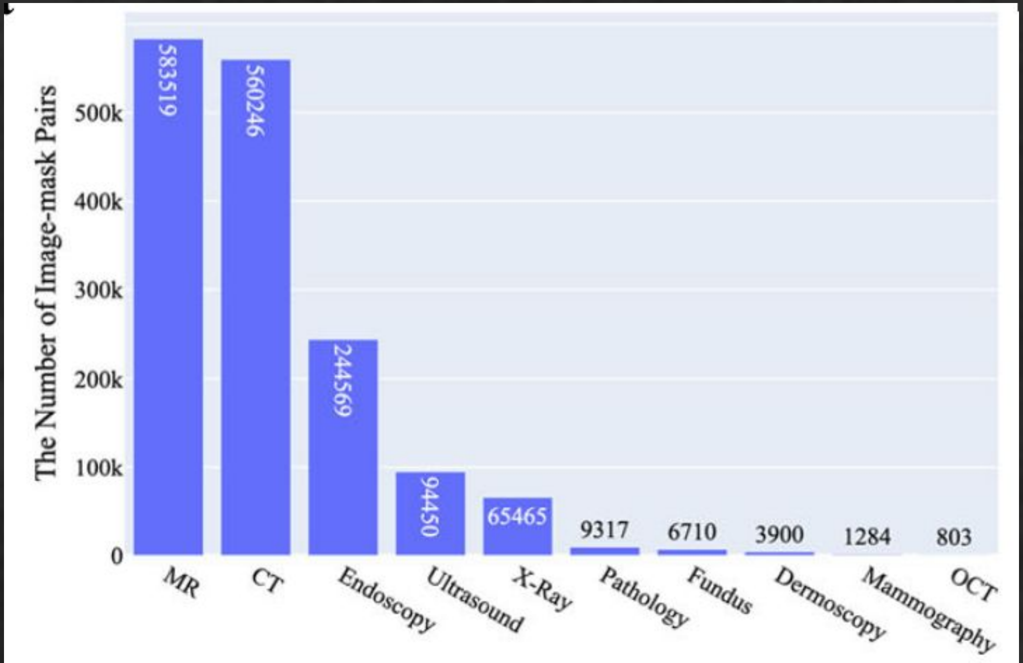
MedSAM is designed as a universal foundation model for medical image segmentation, capable of handling diverse variations in imaging conditions, anatomical structures, and pathologies.



Source : Jun Ma et al.

Large-Scale Dataset

MedSAM was trained on a large-scale dataset containing 1,570,263 medical image-mask pairs across 10 imaging modalities and over 30 cancer types.

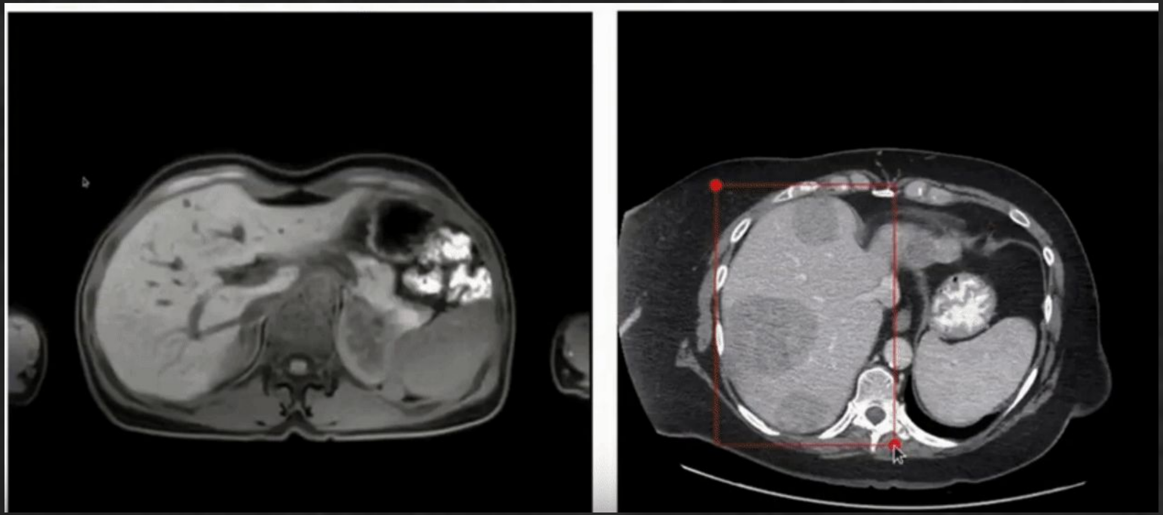


Source : Jun Ma et al.

Fine-Tuning

The large-scale dataset allows MedSAM to learn a rich representation of medical images, capturing a broad spectrum of anatomies and lesions across different modalities.

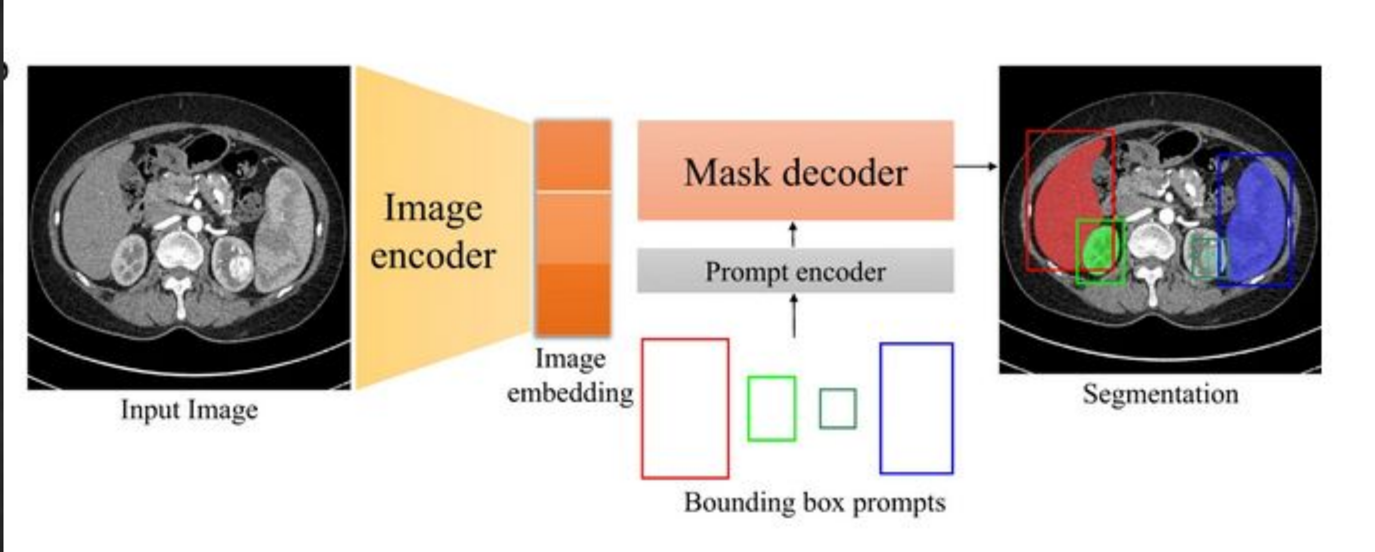
To address SAM's initially unsatisfactory performance in medical image segmentation, in MEDSAM they fine-tuned the model on medical images, using both full fine-tuning and parameter-efficient fine-tuning approaches.



Source : Jun Ma et al.

MedSAM Arch

1. **Promptable Segmentation Model**: MedSAM employs a promptable segmentation approach, using user-provided prompts such as points or bounding boxes to define the segmentation task, enhancing adaptability to specific needs.
2. **Model Architecture**: The network architecture includes an image encoder for converting images into embeddings, a prompt encoder for interpreting user inputs, and a mask decoder that uses cross-attention to integrate these inputs for precise segmentation.



MedSAM Inputs

- **Image Processing**:
  - Input size: 1024×1024×3.
  - Reshaped into 16×16×3 patches, resulting in a 64×64 feature size after encoding.
  - Base ViT Model Used for balancing performance and computational efficiency
    i. Initially pre-trained using masked autoencoder modeling.
    ii. Fully supervised training on the SAM dataset.
- **Prompt Encoding**:
  - Transforms bounding box corners into 256-dimensional vector embeddings.
  - Represents each bounding box with embeddings of the top-left and bottom-right corners.

Mask Decoder Architecture

- **Lightweight Design for Real-Time Interaction**:
  - Two transformer layers fuse image and prompt embeddings.
  - Two transposed convolutional layers upscale the resolution to 256×256.
- **Output Processing**:
  - Sigmoid activation followed by bi-linear interpolations to match the original input size.

Training Protocol & Experimental Setting

Data Pre-processing and Model Development

- **Dataset Composition**: Utilized 1,570,263 medical image-mask pairs.
- **Data Splitting Strategy**:
  - Training (80%), Tuning (10%), and Validation (10%).
  - Special handling for modalities requiring continuity (e.g., CT and MRI segmented at the 3D scan level).

Model Initialization and Training

- **Base Model**: Started with pre-trained SAM model using ViT-Base.
- **Component Configuration**:
  - Fixed the prompt encoder.
  - Trainable parameters for the image encoder and mask decoder were updated.
- **Training Details**:
  - Total parameters: Image Encoder - 89,670,912, Mask Decoder - 4,058,340.
  - Loss function: Unweighted sum of dice loss and cross-entropy loss.
  - Optimization: AdamW optimizer, with a specific focus on learning rates and weight decay.

Training Protocol & Experimental Setting

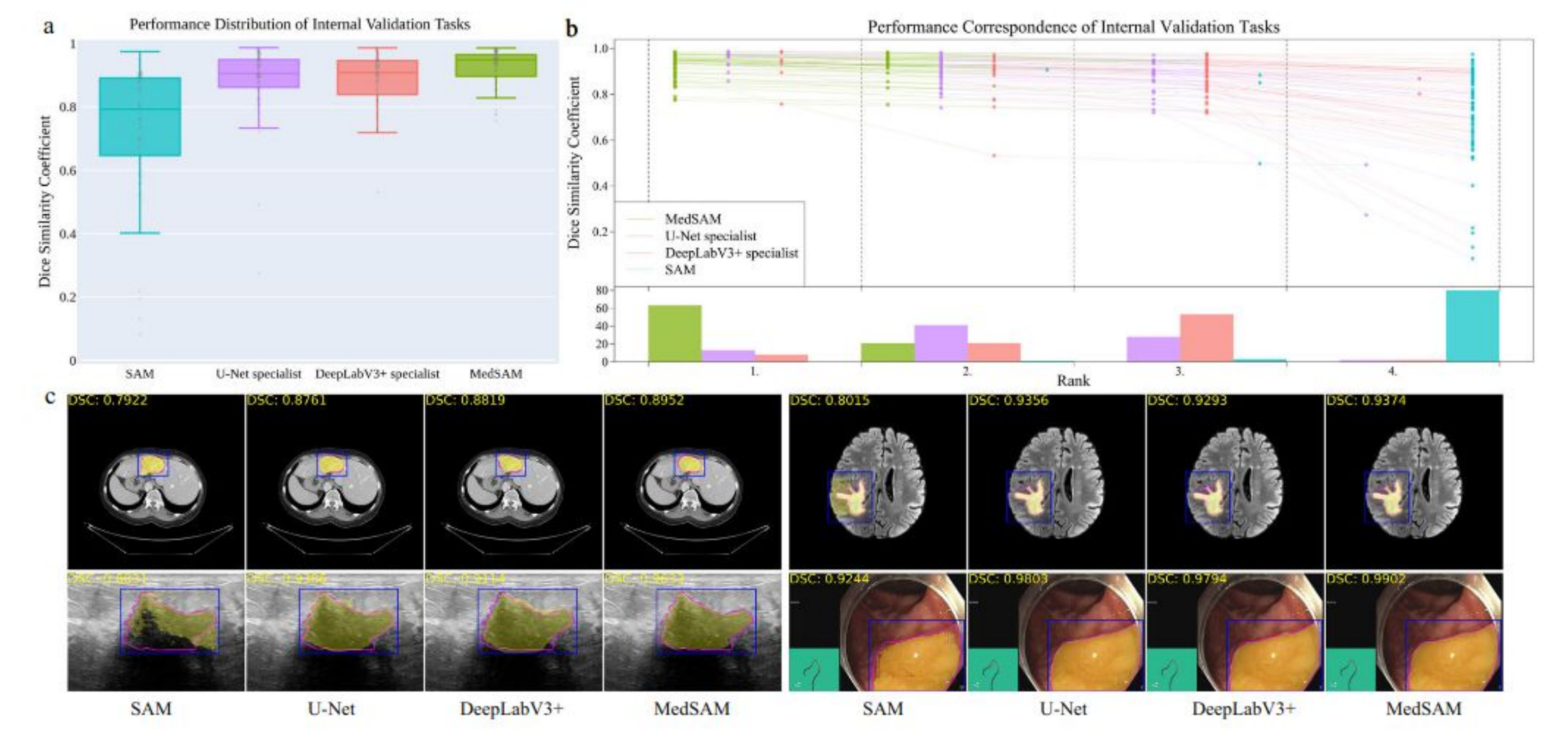
Comparative Performance Evaluation

- **Segmentation Models Compared**: MedSAM vs. state-of-the-art SAM7, U-Net, and DeepLabV3+.
- **Training Modality and Setup**:
  - 10 imaging modalities included.
  - Specialist models trained individually for each modality.
- **Innovative Use of Bounding Boxes**:
  - Transformed into binary masks and used as additional input channels to improve targeting and segmentation accuracy.

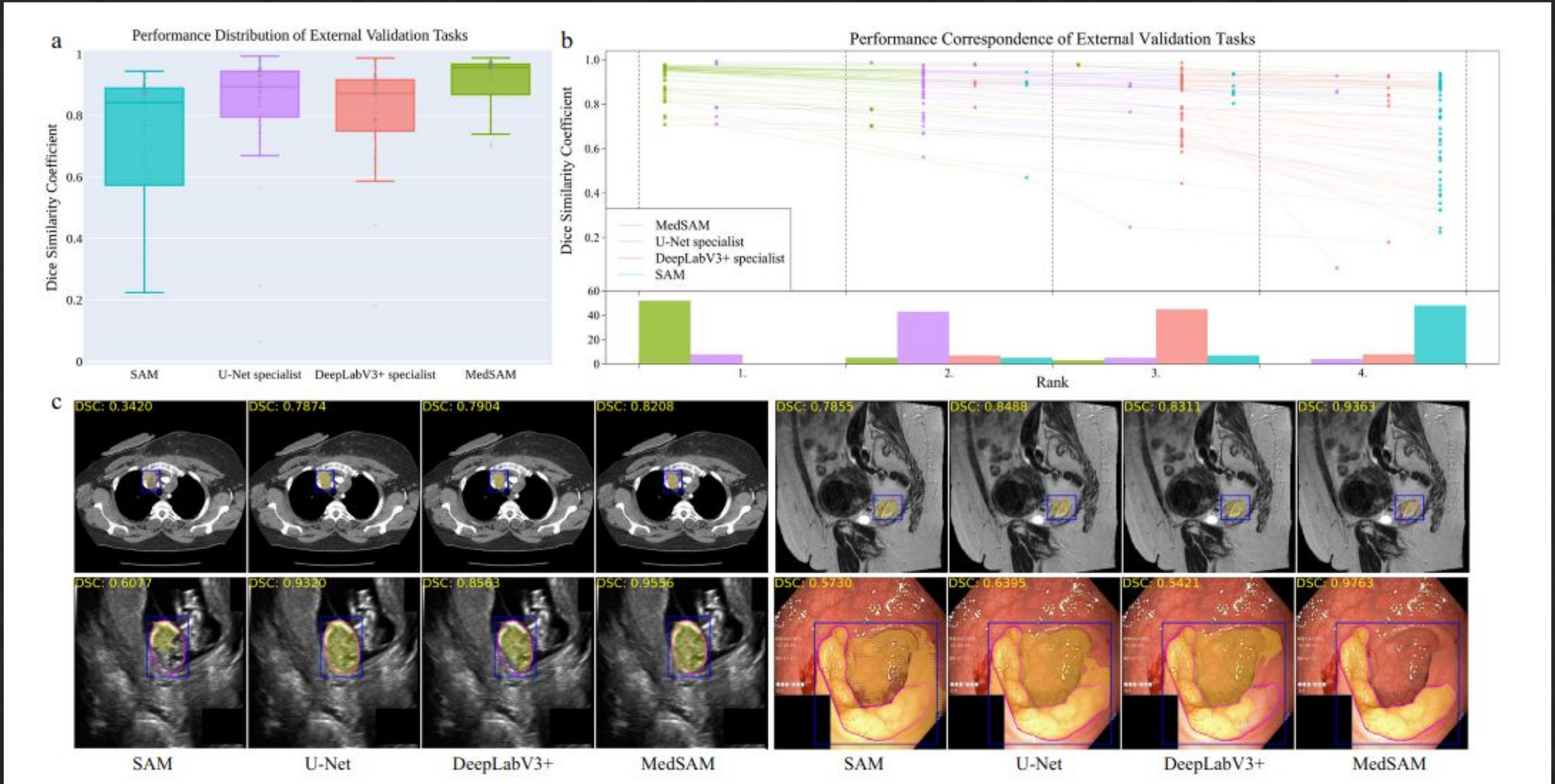
Training Efficiency and Specialization

- **GPU Utilization and Epochs**:
  - MedSAM and specialist models varied in GPU usage and training duration, reflecting optimization for specific tasks.
- **Application Specific Comparisons**:
  - Task-specific U-Net models showcased strong internal validation results but weaker external validation performance.
  - MedSAM demonstrated robust generalization across both validation sets.

Evaluation Results - Internal Validation Set

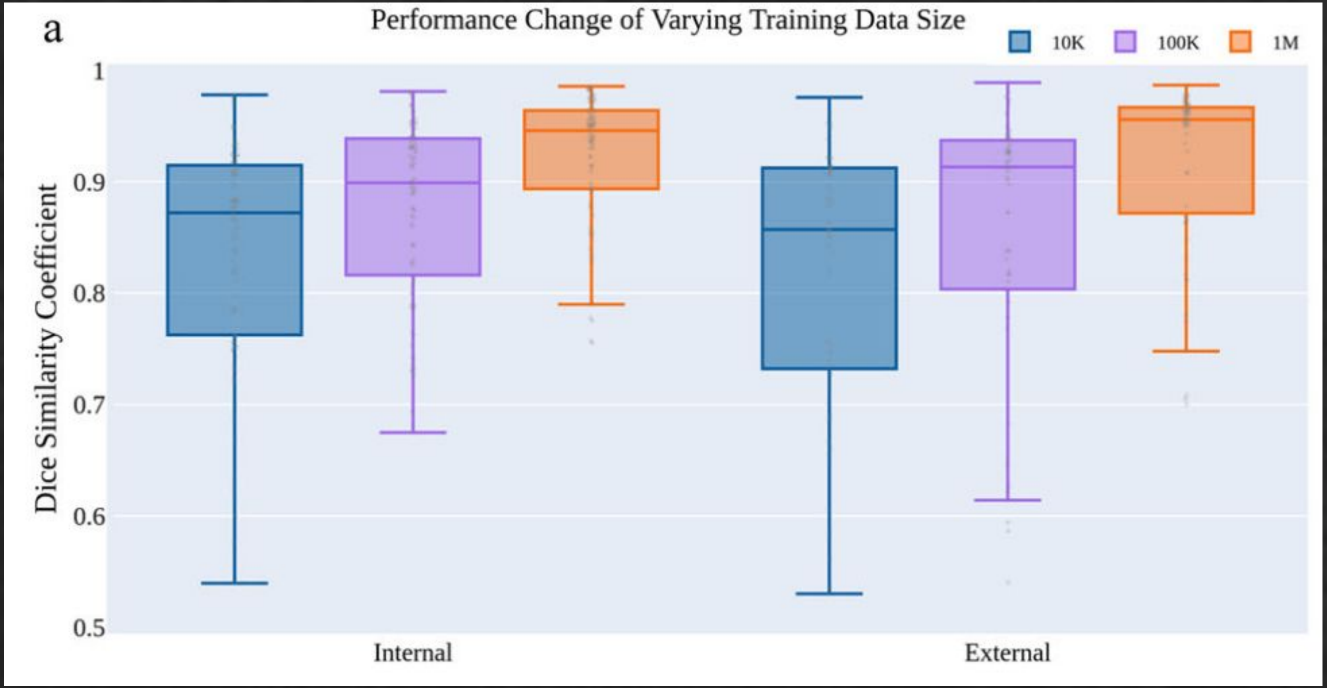


Evaluation Results - External Validation Set



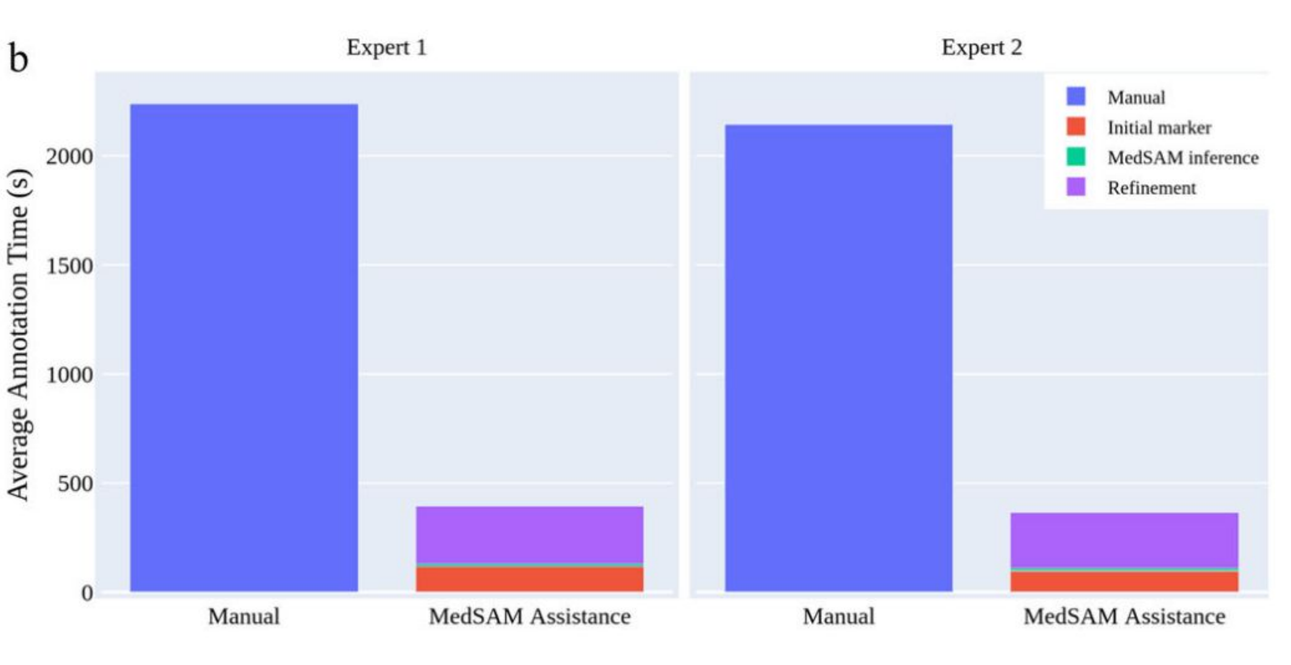
Training Dataset Size Effect

The performance adhered to the scaling rule, where increasing the number of training images significantly improved the performance in both internal and external validation sets



MedSAM & Annotation Efficiency

The results demonstrate that with the assistance of MedSAM, the annotation time is substantially reduced by 82.37% and 82.95% for the two experts, respectively.



Conclusion and Implications

- **Model Generalization**: MedSAM exhibits superior generalization capabilities across diverse medical imaging tasks.

- **Clinical Impact**: Offers a versatile and effective tool for medical image segmentation, adaptable to both common and complex segmentation challenges.

- **Future Directions**: Potential for further refinement of task-specific models to match the robustness of foundational models like MedSAM.

Thank you