Segment Anything in Medical Images

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang

Abstract

Medical image segmentation is a critical component in clinical practice, facilitating accurate diagnosis, treatment planning, and disease monitoring. However, current methods predominantly rely on customized models, which exhibit limited generality across diverse tasks. In this study, we present MedSAM, the inaugural foundation model designed for universal medical image segmentation. Harnessing the power of a meticulously curated dataset comprising over one million images, MedSAM not only outperforms existing state-of-the-art segmentation foundation models, but also exhibits comparable or even superior performance to specialist models. Moreover, MedSAM enables the precise extraction of essential biomarkers for tumor burden quantification. By delivering accurate and efficient segmentation across a wide spectrum of tasks, MedSAM holds significant potential to expedite the evolution of diagnostic tools and the personalization of treatment plans.

INTRODUCTION

Segmentation is a fundamental task in medical imaging analysis, which involves identifying and delineating regions of interest (ROI) in various medical images, such as organs, lesions, and tissues. Accurate segmentation is essential for many clinical applications, including disease diagnosis, treatment planning, and monitoring of disease progression [1], [2]. Manual segmentation has long been the gold standard for delineating anatomical structures and pathological regions, but this process is time-consuming, labor-intensive, and often requires a high degree of expertise. Semi- or fully-automatic segmentation methods can significantly reduce the time and labor required, increase consistency, and enable the analysis of large-scale datasets.

Deep learning-based models have shown great promise in medical image segmentation due to their ability to learn intricate image features and deliver accurate segmentation results across a diverse range of tasks, from segmenting specific anatomical structures to identifying pathological regions [3]. However, a significant limitation of many current medical image segmentation models is their task-specific nature. These models are typically designed and trained for a specific segmentation task, and their performance can degrade significantly when applied to new tasks or different types of imaging data. This lack of generality poses a substantial obstacle to the wider application of these models in clinical practice. In contrast, recent advances in the field of natural image segmentation have witnessed the emergence of segmentation foundation models [4], [5], showcasing remarkable versatility and performance across various segmentation tasks. However, their application to medical image segmentation has been challenging due to the substantial domain gap [6] (Supplementary Related work).

Therefore, there is a growing demand for universal models in medical image segmentation: models that can be trained once and then applied to a wide range of segmentation tasks. Such models would not only exhibit heightened versatility in terms of model capacity, but also potentially lead to more consistent results across different tasks, benefiting from a shared underlying architecture and training process. Motivated by the remarkable generality of the Segment Anything Model (SAM) [4], we introduce MedSAM, the first foundation model for universal medical image segmentation. MedSAM is adapted from the SAM model on an unprecedented scale, with more than one million medical image-mask pairs. We thoroughly evaluate MedSAM through comprehensive experiments on over 70 internal validation tasks and 40 external validation tasks, spanning a variety of anatomical structures, pathological conditions, and medical imaging modalities. Experimental results demonstrate that MedSAM consistently outperforms the state-of-the-art (SOTA) segmentation foundation model, while achieving performance on par with, or even surpassing specialist models. These results highlight the potential of MedSAM as a powerful tool for medical image segmentation.

- *Jun is with Peter Munk Cardiac Centre, University Health Network; Department of Laboratory Medicine and Pathobiology, University of Toronto; Vector Institute, Toronto, Canada*
- *Yuting He is with the Department of Computer Science, Johns Hopkins University, USA.*
- *Feifei Li is with the Department of Cell and Systems Biology, University of Toronto, Canada.*
- *Lin Han is with Tandon School of Engineering, New York University, USA.*
- *Chenyu You is with the Department of Electrical Engineering, Yale University, USA.*
- *Bo Wang (Corresponding Author) is with Peter Munk Cardiac Centre, University Health Network; Department of Laboratory Medicine and Pathobiology and Department of Computer Science, University of Toronto; Vector Institute, Toronto, Canada. E-mail: bowang@vectorinstitute.ai*

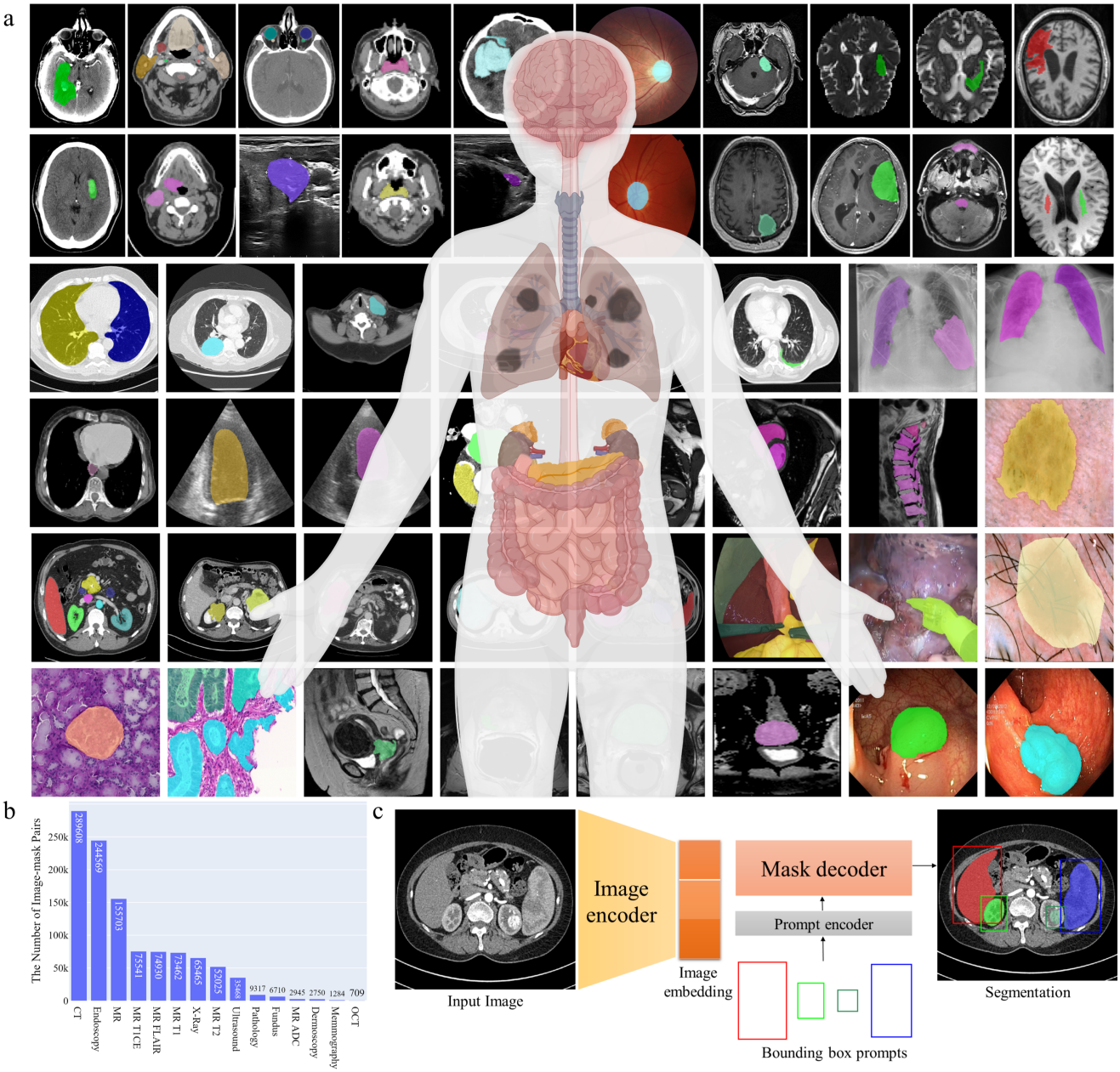

Fig. 1. **MedSAM is trained on a large-scale dataset that can handle diverse segmentation tasks. a,** The dataset covers a variety of anatomical structures, pathological conditions, and medical imaging modalities. **b,** The number of medical image-mask pairs in each modality. **c,** MedSAM is a promptable segmentation method where users can use a bounding box to specify the segmentation target.

Results

MedSAM aims to fulfill the role of a foundation model for universal medical image segmentation. A crucial aspect of constructing such a model is the capacity to accommodate a wide range of variations in imaging conditions, anatomical structures, and pathological conditions. To meet this challenge, we curated a diverse and large-scale medical image segmentation dataset with 1,090,486 medical image-mask pairs, covering 15 imaging modalities, over 30 cancer types, and a multitude of imaging protocols (Fig. 1a, Supplementary Table 1-4). This large-scale dataset allows MedSAM to learn a rich representation of medical images, capturing a broad spectrum of anatomies and lesions across different modalities. Fig. 1b provides an overview of the image distribution across different medical imaging modalities in the dataset, ranked by their total numbers. It is evident that Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and endoscopy are the dominant modalities, reflecting their ubiquity in clinical practice. CT and MRI images provide detailed cross-sectional views of 3D body structures, making them indispensable for non-invasive diagnostic imaging. Endoscopy, albeit more invasive, enables direct visualization of organ interiors, proving invaluable for diagnosing gastrointestinal and urological conditions. Despite the prevalence of these modalities, others such as ultrasound, pathology, fundus, dermoscopy, mammography, and Optical Coherence Tomography (OCT) also hold significant roles in clinical practice. The diversity of these modalities

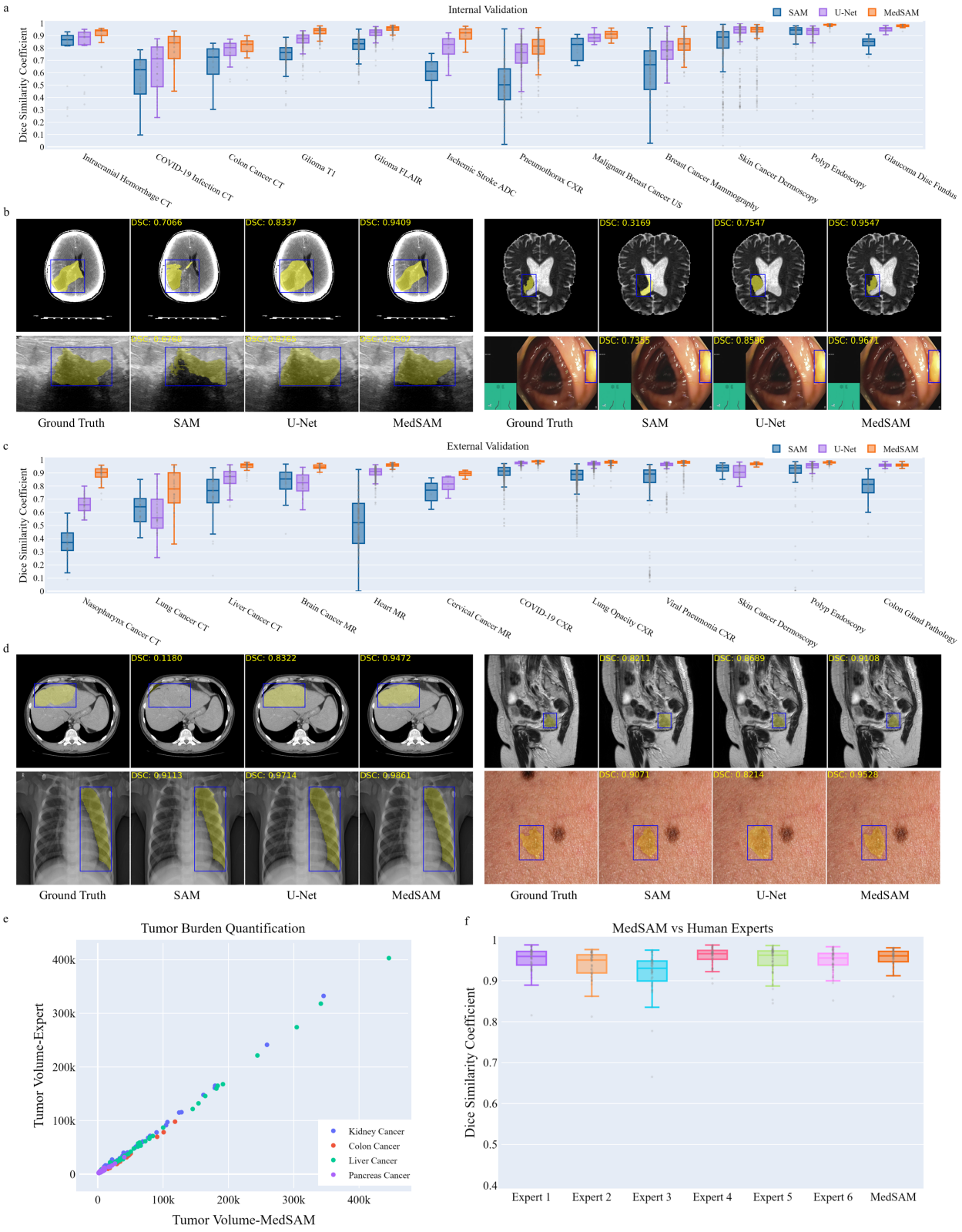

Fig. 2. **MedSAM enables accurate segmentation on a wide range of tasks. a,** Internal validation results of 12 representative segmentation tasks in terms of Dice Similarity Coefficient (DSC) score. The box plots display descriptive statistics across all internal validation cases in each segmentation task, with the median value represented by the horizontal line within the box, the lower and upper quartiles delineating the borders of the box, and the vertical lines indicating the 1.5 interquartile range. **b,** Visualized segmentation examples on the internal validation set show that MedSAM can segment objects with weak boundaries and low contrast. **c,** External validation results of 12 typical segmentation tasks from different modalities. **d,** Visualized segmentation examples on the external validation set show that MedSAM has better generalization ability on unseen datasets or targets. **e,** MedSAM can be used for precise tumor burden quantification. **f,** MedSAM can obtain comparable or even better segmentation accuracy compared to human experts.

and their corresponding segmentation targets underscores the necessity for universal and effective segmentation models capable of handling the unique characteristics associated with each modality.

Another critical consideration is the selection of the appropriate segmentation prompt and network architecture. While the concept of fully automatic segmentation foundation models is enticing, it is fraught with challenges that make it impractical. One of the primary challenges is the variability inherent in segmentation tasks. For example, given a liver cancer CT image, the segmentation task can vary depending on the specific clinical scenario. For instance, one clinician might be interested in segmenting the liver tumor, while another might need to segment the entire liver and surrounding organs. Additionally, the variability in imaging modalities presents another challenge. Modalities such as CT and MR generate 3D images, whereas others like X-Ray and ultrasound yield 2D images. These variabilities in task definition and imaging modalities complicate the design of a fully automatic model capable of accurately anticipating and addressing the diverse requirements of different users.

Considering these challenges, we argue that a more practical approach is to develop a promptable 2D segmentation model. The model can be easily adapted to specific tasks based on user-provided prompts, offering enhanced flexibility and adaptability. It is also able to handle both 2D and 3D images by processing 3D images as a series of 2D slices. Typical user prompts include points and bounding boxes and we show some segmentation examples with the different prompts in Supplementary Fig. 1. It can be found that point-based prompts suffer from ambiguity and require multiple user interventions, whereas the bounding box prompt can clearly specify the ROI with minor user intervention, reducing ambiguity and eliminating trial and error. We follow the network architecture in SAM [4], including an image encoder, a prompt encoder, and a mask decoder (Fig. 1c). The image encoder [7] maps the input image into a high-dimensional image embedding space. The prompt encoder transforms the user-drawn bounding boxes into feature representations via positional encoding [8]. Finally, the mask decoder fuses the image embedding and prompt features using cross-attention [9] (Methods).

We evaluated MedSAM through both internal validation and external validation and compared it to the SOTA segmentation foundation model SAM [4] and specialist U-Net models [3]. The internal validation contained over 70 segmentation tasks (Supplementary Table 5-8, Fig. 2-4), and Fig. 2a shows the Dice Similarity Coefficient (DSC) score of 12 representative segmentation tasks. Overall, SAM obtained inferior performance on most CT, MR, and grey image segmentation tasks although it performed promisingly on some RGB image segmentation tasks, such as skin cancer segmentation (88.8%) in dermoscopy images and polyp (94.1%) segmentation in endoscopy images. This could be attributed to SAM's training on a variety of RGB images, and the fact that many segmentation targets in dermoscopy and endoscopy images are relatively straightforward to segment due to their distinct appearances. Both MedSAM and U-Net outperformed SAM by a large margin on most segmentation tasks ($p < 0.05$), which is expected given their tuning or training on medical image datasets. Compared to the U-Net specialist models, MedSAM still obtained better performance on most tasks. For example, MedSAM achieved median DSC scores of 94.0% (interquartile range (IQR): 91.5-94.9%), 94.4% (IQR: 91.6-95.8%), 81.5% (IQR: 75.1-86.8%), and 98.4% (IQR: 97.9-98.9%) for the segmentation tasks involving intracranial hemorrhage CT, glioma MR T1, pneumothorax CXR, and polyp endoscopy images, respectively, surpassing the performance of the U-Net specialist models by 5%, 6.6%, 5.1%, and 3.6%, respectively. On several RGB image segmentation tasks, such as skin cancer segmentation, the performance between U-Net and MedSAM was comparable (95.1% vs 95.2%). These segmentation targets typically have clear boundaries and good contrasts, making them relatively easy to segment. It's worth noting that U-Net was individually trained for each category (Methods), but MedSAM is a generalist model that was trained only once. Fig. 2b visualizes some segmentation examples of SAM, U-Net, and MedSAM on CT, MR, ultrasound, and endoscopy images. SAM tends to segment the regions with high contrast or clear boundaries, which is prone to under or over-segmentation errors. While the U-Net specialist models offer better segmentation quality, they still struggle with targets with weak boundaries. In contrast, MedSAM can accurately segment a wide range of targets across various imaging conditions, even for objects with weak or missing boundaries (Supplementary Fig.5-7).

The external validation includes over 30 segmentation tasks, all of which are from new datasets or unseen segmentation targets (Supplementary Table 9-11, Fig. 2, 8-9). Fig. 2c shows the DSC score of 12 typical segmentation tasks. SAM continued to exhibit lower performance on most CT and MR segmentation tasks and U-Net specialist models do not consistently outperform SAM (e.g., lung cancer segmentation in CT images (55.8% vs 64.2%)), indicating their limited generalization ability on unseen datasets. In contrast, MedSAM consistently delivers superior performance. For example, MedSAM obtained median DSC scores of 90.3% (IQR: 87.8-93.2%) on the nasopharynx cancer segmentation task, demonstrating 53.3% and 24.5% improvements over SAM and the specialist U-Net, respectively. Significantly, MedSAM also achieved better performance in some unseen modalities (e.g., abdomen T1 Inphase and Outphase), surpassing SAM and specialist U-Net models with improvements by 3-7%. On grey and RGB image segmentation tasks, MedSAM and U-Net specialist models achieved comparable performance in terms of the median DSC score, but MedSAM had fewer outliers. Fig. 2d presents four segmentation examples for qualitative evaluation, revealing that while all the methods have the ability to handle simple segmentation targets, MedSAM performs better at segmenting challenging targets, such as liver cancer in CT images and cervical cancer in MR images (Supplementary Fig. 10). Furthermore, we conducted a visualization and comparative analysis of the saliency maps for image embeddings between SAM and MedSAM (Supplementary Fig. 11). Notably, MedSAM's features exhibited a greater abundance of semantic information, specifically pertaining to highly relevant anatomical structures. Altogether, these results demonstrate that MedSAM has strong generalization abilities across new datasets.

Beyond its broad applicability, we further show that MedSAM facilitates the precise quantification of tumor burden, a critical biomarker in oncology practice [10] (Fig. 2e). Specifically, we calculated the tumor volumes for kidney, colon, liver, and pancreatic cancers using MedSAM segmentation results and compared these with volumes derived from expert segmentation. The tumor volumes obtained from MedSAM and expert evaluations exhibited a high Pearson correlation ($r = 0.99$), underscoring that MedSAM's segmentation results can be effectively utilized for accurate tumor burden quantification. Lastly, we compared MedSAM's performance to that of six human experts in prostate segmentation (Methods). MedSAM's performance was found to be on par with four human experts and even surpassed that of two experts, highlighting its potential as a robust tool for medical image segmentation in clinical practice.

DISCUSSION

We introduce MedSAM, a deep learning-powered foundation model designed for the segmentation of a wide array of anatomical structures and lesions across diverse medical imaging modalities. MedSAM is trained on a meticulously assembled large-scale dataset comprising over one million medical image-mask pairs. Its promptable configuration strikes an optimal balance between automation and customization, rendering MedSAM a versatile tool for universal medical image segmentation.

Through comprehensive evaluations encompassing both internal and external validation, MedSAM has demonstrated substantial capabilities in segmenting a diverse array of targets and robust generalization abilities to manage new data and tasks. Its performance not only significantly exceeds that of existing state-of-the-art segmentation foundation models, but also rivals or even surpasses specialist models. By providing precise delineation of anatomical structures and pathological regions, MedSAM facilitates the computation of various quantitative measures that serve as biomarkers. For instance, in the field of oncology, MedSAM could play a crucial role in generating accurate tumor segmentation results, enabling subsequent calculations of tumor volume, which is a critical biomarker for assessing disease progression and response to treatment.

While MedSAM boasts strong capabilities, it does present certain limitations. One such limitation is the modality imbalance in the training set, with CT, MRI, and endoscopy images dominating the dataset. This could potentially impact the model's performance on less-represented modalities, such as mammography. Another limitation is its difficulty in the segmentation of vessel-like branching structures because the bounding box prompt can be ambiguous in this setting. For example, arteries and veins share the same bounding box in eye fundus images. However, these limitations do not diminish MedSAM's utility. Since MedSAM has learned rich and representative medical image features from the large-scale training set, it can be fine-tuned to effectively segment new tasks from less-represented modalities or intricate structures like vessels.

In conclusion, this study highlights the feasibility of constructing a single foundation model capable of managing a multitude of segmentation tasks, thereby eliminating the need for task-specific models. MedSAM, as the inaugural foundation model in medical image segmentation, holds great potential to accelerate the advancement of new diagnostic and therapeutic tools, and ultimately contribute to improved patient care [11].

METHODS

Study design

Segmentation is an essential step in many medical image-based clinical analysis tasks. For instance, in brain tumor imaging, segmentation of MR images can help identify the location, size, and type of tumor, which are crucial for surgical planning and prognosis [12]. In cardiac imaging, segmentation of structures such as the left ventricle in echocardiograms or MRIs is essential for assessing cardiac function and diagnosing conditions like heart failure [13]. In pulmonary imaging, segmentation of lung fields in chest X-rays or CT images is crucial for diagnosing and monitoring conditions like chronic obstructive pulmonary disease (COPD) and COVID-19 [14]. Over the past few decades, the field of medical image segmentation has witnessed the development of numerous methodologies [15]. However, a significant limitation of many existing approaches is their task-specific and dataset-specific nature, rendering them incapable of generalizing to novel datasets and targets. This limitation hinders their widespread application in clinical practice.

Recent advancements in the field of deep learning, particularly with the introduction of foundation models such as Segment Anything Model (SAM) [4], have shown great potential in addressing the challenge of generalization in medical image segmentation. Foundation models leverage vast amounts of training data and powerful architectures to capture complex patterns and relationships within images. To investigate SAM's applicability in the medical domain, we conducted an experiment using SAM to segment a representative abdominal CT image. SAM offers three primary segmentation modes: fully automatic segmentation, bounding box mode, and point mode. Although text prompts were incorporated into SAM's training pipeline, it's important to note that this is a proof-of-concept and not publicly available in SAM's official repository. Supplementary Fig. 1 illustrates the results obtained from the three segmentation modes. These results were generated using the out-of-the-box online demo, accessible at https://segment-anything.com/demo. In the segment-everything mode, SAM divided the entire image into six distinct regions based on image intensity (Supplementary Fig. 1b). However, the utility of such segmentation results was limited due to two primary reasons. First, the segmented regions lacked semantic labels, making it challenging to interpret the specific anatomical structures. Second, in clinical scenarios,

healthcare professionals predominantly focus on meaningful regions of interest (ROIs), such as the liver, kidneys, spleen, and lesions.

On the other hand, the bounding box-based segmentation mode demonstrated promising results, especially for the right kidney, achieved by providing the upper-left and bottom-right points (Supplementary Fig. 2c). For the point-based segmentation mode (Supplementary Fig. 1d), we initially supplied a single foreground point representing the center of the right kidney. However, SAM over-segmented the entire abdomen. To rectify this, we introduced a background point within the over-segmented regions. This adjustment resulted in the segmentation mask shrinking to encompass only the liver and right kidney. Finally, by adding another background point on the liver, we obtained the desired kidney segmentation.

To summarize, when employing SAM for medical image segmentation, the segment-everything mode often produces partitions that lack practical utility, while the point-based mode can be ambiguous and necessitates multiple iterations for prediction and correction. Conversely, the bounding box-based mode offers a distinct advantage by precisely defining the region of interest (ROI) and consistently yielding reasonable segmentation results, eliminating the need for iterative trial and error. However, despite its promising potential, recent studies have demonstrated that SAM has encountered challenges in delivering satisfactory segmentation results across various medical image segmentation tasks. In light of these limitations, the objective of this study is to develop a robust segmentation foundation model capable of effectively addressing a wide range of segmentation targets and diverse imaging modalities. The subsequent subsections provide a comprehensive overview of the key aspects encompassing the training and (internal and external) validation sets, network architecture, training protocol, and the comparison with state-of-the-art baselines.

Dataset curation and pre-processing

We curated a comprehensive dataset by collating images from publicly available medical image segmentation datasets, which were obtained from various sources across the internet. These sources include The Cancer Imaging Archive (TCIA) at https://www.cancerimagingarchive.net/, Kaggle at https://www.kaggle.com/, Grand-Challenge at https://grand-challenge.org/challenges/, Scientific Data at https://www.nature.com/sdata/, CodaLab at https://codalab.lisn.upsaclay.fr/, and segmentation challenges within the Medical Image Computing and Computer Assisted Intervention Society (MICCAI) at http://www.miccai.org/. The complete list of datasets utilized is presented in Supplementary Table 1-4.

The original 3D datasets consisted of Computed Tomography (CT) and Magnetic Resonance (MR) images in DICOM, nrrd, or mhd formats. To ensure uniformity and compatibility with developing medical image deep learning models, we converted the images to the widely used NifTI format. Additionally, grayscale images (such as X-Ray and Ultrasound) as well as RGB images (including endoscopy, dermoscopy, fundus, and pathology images), were converted to the png format. Several exclusive criteria are applied to improve the dataset quality and consistency, including incomplete images and segmentation targets with branching structures, inaccurate annotations, and tiny volumes. Notably, image intensities varied significantly across different modalities. For instance, CT images had intensity values ranging from -2000 to 2000, while MR images exhibited a range of 0 to 3000. In endoscopy and ultrasound images, intensity values typically spanned from 0 to 255. To facilitate stable training, we performed intensity normalization across all images, ensuring they shared the same intensity range.

For CT images, we initially normalized the Hounsfield units using typical window width and level values, as outlined in https://radiopaedia.org/articles/windowing-ct. Subsequently, the intensity values were rescaled to the range of [0, 255]. For MR, X-Ray, ultrasound, mammography, and Optical Coherence Tomography (OCT) images, we clipped the intensity values to the range between the 0.95th and 99.5th percentiles before rescaling them to the range of [0, 255]. Regarding RGB images (e.g., endoscopy, dermoscopy, fundus, and pathology images), if they were already within the expected intensity range of [0, 255], their intensities remained unchanged. However, if they fell outside this range, we utilized max-min normalization to rescale the intensity values to [0, 255]. Finally, to meet the model's input requirements, all images were resized to a uniform size of $1024 \times 1024 \times 3$. In the case of whole-slide pathology images, patches were extracted using a sliding window approach. As for 3D CT and MR images, each 2D slice was resized to 1024×1024, and the channel was repeated three times to maintain consistency. The remaining 2D images were directly resized to $1024 \times 1024 \times 3$. Bi-cubic interpolation was used for resizing images, while nearest-neighbor interpolation was applied for resizing masks to preserve their precise boundaries and avoid introducing unwanted artifacts. These standardization procedures ensured uniformity and compatibility across all images and facilitated seamless integration into the subsequent stages of the model training and evaluation pipeline.

Network architecture

The network utilized in this study was built on transformer architecture [9], which has demonstrated remarkable effectiveness in various domains such as natural language processing [16] and image recognition tasks [7]. Specifically, the network incorporated a vision transformer (ViT)-based image encoder responsible for extracting image features, a prompt encoder for integrating user interactions (bounding boxes), and a mask decoder that generated segmentation results and confidence scores using the image embedding, prompt embedding, and output token.

To strike a balance between segmentation performance and computational efficiency, we employed the base ViT model as the image encoder since extensive evaluation indicated that larger ViT models, such as ViT Large and ViT Huge, offered

only marginal accuracy improvements [4] while significantly increasing computational demands. Specifically, the base ViT model consists of 12 transformer layers [9], with each block comprising a multi-head self-attention block and a Multilayer Perceptron (MLP) block incorporating layer normalization [17]. Pre-training was performed using masked auto-encoder modeling [18], followed by fully supervised training on the SAM dataset [4]. The input image ($1024 \times 1024 \times 3$) was reshaped into a sequence of flattened 2D patches with the size $16 \times 16 \times 3$, yielding a feature size in image embedding of 64×64 after passing through the image encoder, which is $16\times$ downscaled. The prompt encoders mapped the corner point of the bounding box prompt to 256-dimensional vectorial embeddings [8]. In particular, each bounding box was represented by an embedding pair of the top-left corner point and the bottom-right corner point. To facilitate real-time user interactions once the image embedding had been computed, a lightweight mask decoder architecture was employed. It comprised two transformer layers [9] for fusing the image embedding and prompt encoding, and two transposed convolutional layers to enhance the embedding resolution to 256×256. Subsequently, the embedding underwent sigmoid activation, followed by bi-linear interpolations to match the input size.

Training protocol and experimental setting

During data pre-processing, we obtained 1,090,486 medical image-mask pairs for model development (do not include the external validation sets Table 1-4). For internal validation, we randomly split the dataset into 80%, 10%, and 10% as training, tuning, and validation, respectively. This setup allowed us to monitor the model's performance on the tuning set and adjust its parameters during training to prevent overfitting. For external validation, we used hold-out datasets that were not seen by the model during training. These datasets provide a stringent test of the model's generalization ability, as they represent new patients, imaging conditions, and potentially new segmentation tasks that the model has not encountered before. By evaluating the performance of MedSAM on these unseen datasets, we can gain a realistic understanding of how MedSAM is likely to perform in real-world clinical settings, where it will need to handle a wide range of variability and unpredictability in the data. The training and validation are independent.

The model was initialized with the pre-trained SAM model with the ViT-Base model. We fix the prompt encoder since it can already encode the bounding box prompt. All the trainable parameters in the image encoder and mask decoder were updated during training. Specifically, the number of trainable parameters for the image encoder and mask decoder are 89,670,912 and 4,058,340, respectively. The bounding box prompt was simulated from the ground-truth mask with a random perturbation of 0-20 pixels. The loss function is the unweighted sum between Dice loss and cross-entropy loss, which has been proven to be robust in various segmentation tasks [3]. Specifically, let S, G denote the segmentation result and ground truth, respectively. s_i, g_i denote the predicted segmentation and ground truth of voxel i, respectively. N is the number of voxels in the image I. Cross-entropy loss is defined by

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N} g_i \log s_i,$$

and dice loss is defined by

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} g_i s_i}{\sum_{i=1}^{N}(g_i)^2 + \sum_{i=1}^{N}(s_i)^2}.$$

The final loss L is defined by

$$L = L_{CE} + L_{Dice}.$$

The network was optimized by AdamW [19] optimizer (β_1 = 0.9, β_2 = 0.999) with an initial learning rate of 1e-4 and a weight decay of 0.01. The batch size was 160 and data augmentation was not used. The model was trained on 20 A100 (80G) GPUs with 100 epochs and the last checkpoint was selected as the final model.

Furthermore, to thoroughly evaluate the performance of MedSAM, we conducted comparative analyses against both the state-of-the-art segmentation foundation model SAM [4] and specialist models. Specifically, we divided the training images into four categories: CT images, MR images, grey images (including chest X-Ray (CXR), ultrasound, mammography, and OCT images), and RGB images (including pathology, endoscopy, and dermoscopy images). For each category, we trained a specialist model based on nnU-Net, which achieved SOTA performance on many segmentation tasks [3]. To generate the training data for U-Net, we cropped the images and corresponding masks inside the bounding boxes. For a fair comparison, both MedSAM and U-Net specialist models were trained on the same data split. The primary difference was in the training method: MedSAM underwent training once on the entire training set, while U-Net specialist models were trained separately on each subset corresponding to one category.

In addition to the comparative analyses against SAM and U-Net specialist models, we further assessed MedSAM's performance by comparing it to six experts on a prostate MR image segmentation dataset (52 cases). For each case, the six experts provided their respective segmentation results, and the ground truth was determined based on majority voting. We calculated the DSC scores for each case and expert, subsequently comparing them to MedSAM's results.

Evaluation metrics

We follow the recommendations in Metric Reload [20] and use Dice Similarity Coefficient and Normalized Surface Distance (NSD) to quantitatively evaluate the segmentation results. DSC is a region-based segmentation metric, aiming to evaluate the region overlap between ground truth and segmentation results, which is defined by

$$DSC(G,S) = \frac{2|G \cap S|}{|G| + |S|},$$

NSD is a boundary-based metric, aiming to evaluate the boundary consensus between ground truth and segmentation results at a given tolerance, which is defined by

$$NSD(G,S) = \frac{|\partial G \cap B_{\partial S}^{(\tau)}| + |\partial S \cap B_{\partial G}^{(\tau)}|}{|\partial G| + |\partial S|},$$

where $B_{\partial G}^{(\tau)} = \{x \in R^3 \,|\, \exists \tilde{x} \in \partial G,\, ||x - \tilde{x}|| \leq \tau\}$, $B_{\partial S}^{(\tau)} = \{x \in R^3 \,|\, \exists \tilde{x} \in \partial S,\, ||x - \tilde{x}|| \leq \tau\}$ denote the border region of the ground truth and the segmentation surface at tolerance τ, respectively. In this paper, we set the tolerance τ as 2.

Statistical analysis

To statistically analyze and compare the performance of the aforementioned three methods (MedSAM, SAM, and specialist models), we employed the Wilcoxon signed-rank test. This non-parametric test is well-suited for comparing paired samples and is particularly useful when the data does not meet the assumptions of normal distribution. This analysis allowed us to determine if any method demonstrated statistically superior segmentation performance compared to the others, providing valuable insights into the comparative effectiveness of the three evaluated methods: SAM, U-Net specialist models, and MedSAM. The Wilcoxon signed-rank test results are marked on the DSC and NSD score tables (Supplementary Table 5-11).

Data availability

All the datasets in this study are from public datasets. The download links are provided in Supplementary Table 12.

Code availability

All code was implemented in Python (3.10) using Pytorch (2.0) as the base deep learning framework. We also used several python packages for data analysis and results visualization, including SimpleITK (2.2.1), nibabel (5.1.0), torchvision (0.15.2), numpy (1.24.3), scikit-image (0.20.0), opencvpython (4.7.0), scipy (1.10.1), and pandas (2.0.2), matplotlib (3.7.1), and plotly (5.15.0). Biorender was used to create Fig. 1a. The training script, inference script, and trained model have been publicly available at https://github.com/bowang-lab/MedSAM.

Acknowledgements

The authors of this paper highly appreciate all the data owners for providing public medical images to the community. We also thank Meta AI for making the source code of segment anything publicly available to the community.

REFERENCES

[1] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[2] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley *et al.*, "Video-based ai for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, pp. 252–256, 2020.

[3] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[4] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[5] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Gao, and Y. J. Lee, "Segment everything everywhere all at once," *arXiv preprint arXiv:2304.06718*, 2023.

[6] 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?" *arXiv preprint arXiv:2304.14660*, 2023.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[8] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7537–7547, 2020.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural Information Processing Systems*, vol. 30, 2017.

[10] E. K. Abdalla, R. Adam, A. J. Bilchik, D. Jaeck, J.-N. Vauthey, and D. Mahvi, "Improving resectability of hepatic colorectal metastases: expert consensus statement," *Annals of Surgical Oncology*, vol. 13, pp. 1271–1280, 2006.

[11] K. Bera, N. Braman, A. Gupta, V. Velcheti, and A. Madabhushi, "Predicting cancer outcomes with radiomics and artificial intelligence in radiology," *Nature Reviews Clinical Oncology*, vol. 19, no. 2, pp. 132–146, 2022.

[12] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.

[13] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martín-Isla, A. Sojoudi, P. M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma, M. Parreño, A. Albiol, F. Kong, S. C. Shadden, J. C. Acero, V. Sundaresan, M. Saber, M. Elattar, H. Li, B. Menze, F. Khader, C. Haarburger, C. M. Scannell, M. Veta, A. Carscadden, K. Punithakumar, X. Liu, S. A. Tsaftaris, X. Huang, X. Yang, L. Li, X. Zhuang, D. Viladés, M. L. Descalzo, A. Guala, L. La Mura, M. G. Friedrich, R. Garg, J. Lebel, F. Henriques, M. Karakas, E. Çavuş, S. E. Petersen, S. Escalera, S. Seguí, J. F. Rodríguez-Palomares, and K. Lekadir, "Multi-centre, multi-vendor and multi-disease cardiac segmentation: The m&ms challenge," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3543–3554, 2021.

[14] K. Zhang, X. Liu, J. Shen, Z. Li, Y. Sang, X. Wu, Y. Zha, W. Liang, C. Wang, K. Wang *et al.*, "Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography," *Cell*, vol. 181, no. 6, pp. 1423–1433, 2020.

[15] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, 2017.

[16] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[17] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[18] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[19] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[20] L. Maier-Hein, A. Reinke, P. Godau, M. D. Tizabi, F. Büttner *et al.*, "Metrics reloaded: Pitfalls and recommendations for image analysis validation," *arXiv preprint arXiv:2206.01653*, 2022.