scientific reports

Check for updates

OPEN

An efficient segment anything model for the segmentation of medical images

Guanliang Dong[1], Zhangquan Wang[2]✉, Yourong Chen[2], Yuliang Sun[2], Hongbo Song[2], Liyuan Liu[3] & Haidong Cui[4]

This paper introduces the efficient medical-images-aimed segment anything model (EMedSAM), addressing the high computational demands and limited adaptability of using SAM for medical image segmentation tasks. We present a novel, compact image encoder, DD-TinyViT, designed to enhance segmentation efficiency through an innovative parameter tuning method called med-adapter. The lightweight DD-TinyViT encoder is derived from the well-known ViT-H using a decoupled distillation approach.The segmentation and recognition capabilities of EMedSAM for specific structures are improved by med-adapter, which dynamically adjusts the model parameters specifically for medical imaging. We conducted extensive testing on EMedSAM using the public FLARE 2022 dataset and datasets from the First Hospital of Zhejiang University School of Medicine. The results demonstrate that our model outperforms existing state-of-the-art models in both multi-organ and lung segmentation tasks.

Keywords Segment anything model, Medical image, Lightweight encoder, Decoupled distillation, Fine tuning adapter

Medical image analysis involves the use of various algorithms and techniques to process and analyze medical imaging data, aiming to enhance the accuracy of physicians' diagnoses and treatments[1]. Among these techniques, medical image segmentation is particularly critical, as it isolates pathological regions and extracts essential information, thereby enabling physicians to better analyze, interpret, and understand the images. For instance, in tumor diagnosis, precise image segmentation helps determine the tumor's size, location, and morphology[2]. Similarly, in diagnosing cardiovascular diseases, accurate segmentation of cardiac structures facilitates improved assessments of heart function and blood flow dynamics[3]. Consequently, the field of medical image segmentation has garnered significant attention from both medical and academic communities.

Early medical image segmentation methods mainly depended on traditional mathematical approaches like thresholding, region growing, edge detection, and clustering. Although these methods are straightforward to implement, they rely significantly on domain-specific prior knowledge and generally have poor generalization capabilities[4]. With the advancement of artificial intelligence (AI) over recent decades, researchers have developed various image segmentation techniques based on machine learning, including decision trees, support vector machines, and random forests. These methods have improved segmentation accuracy; however, they tend to overfit, leading to limited robustness and inconsistent performance.

Currently, the leading methods for medical image segmentation employ deep learning techniques, such as convolutional neural networks (CNNs), fully convolutional networks (FCNs), and U-Net. These approaches achieve rapid and highly accurate segmentation by automatically extracting features from images. Nonetheless, CNNs face challenges related to the depth of their network architecture and the availability of data. FCNs, while effective, may lose critical edge details during the pooling process, which can adversely affect their performance[5]. U-Net, known for requiring fewer samples and providing rapid segmentation capabilities, struggles when target regions constitute a minor fraction of the entire image. The abundance of extracted features in U-Net can also lead to overfitting issues[6]. Moreover, these neural network models are often tailored to specific patterns and targets, resulting in limited generalization capabilities.

[1]School of Information Engineering, Huzhou University, Huzhou 313000, China. [2]College of Information Science and Technology, Zhejiang Shuren University, Hangzhou 310015, China. [3]Department of Decision and System Sciences, Saint Joseph's University, Philadelphia 19131, USA. [4]Department of Breast Surgery, First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310003, China. ✉email: zqwang@zjsru.edu.cn

In practical therapeutic processes, physicians often rely on multiple types of imaging data for disease diagnosis, including X-rays, computed tomography (CT), ultrasound, and magnetic resonance imaging (MRI). However, most medical image segmentation methods typically focus on specific organs or imaging modalities. A critical issue with these methods is their poor generalization capability across different types of data[7]. The segment anything model (SAM), proposed by Meta AI and Facebook AI Research in 2023, is a versatile image segmentation model that employs deep learning techniques to automatically adapt to various image segmentation tasks[8]. Since there is no need to redesign or adjust the model architecture for each specific task, SAM exhibits strong zero-shot generalization capabilities. As a result, it has garnered increasing research interest in the field of medical image segmentation[9].

The performance of the basic SAM model was evaluated on common medical datasets such as FLARE 2022[10,11]. The results indicated that the model performed well on targets with clear boundaries but struggled with targets that have indistinct boundaries and continuous structures, such as blood vessels. Additionally, SAM's segmentation capabilities were assessed using CT medical images, demonstrating its effectiveness[12]. However, it should be noted that the basic SAM model was not fine-tuned in the studies[10–12], resulting in a performance gap compared to state-of-the-art segmentation methods. SAM's performance was enhanced by incorporating medical knowledge through the insertion of adapter modules at specific positions within the model[13]. Moreover, a low-rank adaptation (LoRA) strategy was utilized to fine-tune the SAM encoder, developing an effective solution for medical image segmentation[14]. The results suggested that the fine-tuned SAM model could achieve performance comparable to current state-of-the-art methods. However, these studies did not address the model's limitations in processing speed, especially in rapidly segmenting large volumes of images from CT and other imaging modalities, which is critical for prompt diagnosis.

In summary, while the SAM model demonstrates strong generalization capabilities, it faces dual challenges in terms of segmentation efficiency and accuracy, particularly when dealing with highly complex and real-time medical images. Therefore, this paper introduces the efficient medical-images-aimed segment anything model (EMedSAM), which enhances performance through a novel combination of a parameter adaptation mechanism and a lightweight image encoder. First, we present med-adapter, a meticulously designed parameter adaptation mechanism strategically placed between the encoder and decoder, paired with specific training strategies to boost the model's adaptability. Additionally, we incorporate a lightweight image encoder based on the TinyViT architecture, named DD-TinyViT. This encoder utilizes decoupled distillation techniques to effectively transfer learned knowledge from the ViT-H model, achieving efficient and accurate image segmentation. The seamless integration of these components not only enhances segmentation speed but also ensures high segmentation accuracy. The main contributions of this paper are as follows:

1. To enhance the flexibility of the SAM model and better adapt it to the characteristics of medical images, we propose an innovative parameter adaptation strategy: med-adapter. This novel mechanism dynamically fine-tunes model parameters, significantly improving the accuracy and versatility of the segmentation process.
2. To significantly enhance processing speed, we introduce a novel lightweight image encoder, DD-TinyViT, crafted through a process known as decoupled distillation. This development meets the urgent demands of clinical practice, facilitating rapid diagnosis and treatment.
3. We enhance the performance and generalization ability of medical image segmentation. Our model sets a new benchmark and performs excellently on both established public datasets and those from the First Hospital of Zhejiang University School of Medicine, demonstrating its effectiveness in practical applications.

Related work

Medical image segmentation based on deep learning

In recent years, a significant number of researchers integrated prevalent deep learning techniques into the realm of medical image segmentation. Several studies delved into the conventional methodology of CNNs, contributing notable advancements in segmenting medical images. For example, Paluru et al.[15] introduced Anam-Net, a streamlined CNN that utilizes altered depth embedding to detect anomalies in COVID-19 chest CT scans. This model achieves performance comparable to U-Net and its variants while maintaining a lightweight structure. Wang et al.[16] introduced DeepIGeoS, an interactive segmentation method leveraging CNN technology. This method uses one CNN for initial automatic segmentation and another CNN to incorporate user interaction with the initial segmentation, enabling 2D placenta segmentation and 3D brain tumor segmentation and thus enhancing segmentation accuracy. However, the CNN models lack spatial awareness when segmenting small targets and are prone to saturation as the network depth increases, leading to performance degradation[15,16].

Following this trend, some researchers explored the application of FCNs for medical image segmentation. FCNs replace the fully connected layer with a fully convolutional layer and allow the output to have the same spatial resolution as the input image. For example, Wang et al.[17] proposed a boundary-sensitive representation model based on multi-task learning. This model employs an FCN with boundary-sensitive representation for precise prostate segmentation in CT images, and can more accurately represent semantic boundary information. However, the FCN model does not have a clear mechanism to capture global context information around pixels. When there is overlap between objects, complex backgrounds, fine boundaries, and insufficient detail information, it will lead to segmentation inaccuracies.

Some researchers utilized U-Net for medical image segmentation. The U-Net features an encoder-decoder structure, incorporates skip connections, and enables pixel classification for effective multi-scale analysis. For example, Sun et al.[18] introduced a multi-scale contextual attention-based network (MSCA-Net). By adeptly fusing the encoder and decoder, a global-local channel space attention module was devised to capture contextual information. Isensee et al.[19] unveiled an out-of-the-box nnU-Net model that excels beyond many current techniques

in international biomedical segmentation competitions. Liu et al. presented a deep U-Net model to solve the edge leakage problem. The model employs a super-pixel region merging process for edge enhancement and a bilateral filtering model to eliminate external noise[20]. Although U-Net is effective in medical image segmentation, it struggles with global contextual information. IT also exhibits challenges in handling complex scene images, and faces difficulties in addressing long-range dependencies. It needs necessitate more labeled data and is less suited for complex and multi-faceted segmentation tasks.

Therefore, the attention has shifted towards employing Transformers for medical image segmentation, which are particularly adept at capturing long-distance dependencies and global context information. Thus it enhances the model's effectiveness, especially with large-scale medical image datasets. Transformers excel in understanding inter-pixel relationships, which is instrumental in improving segmentation accuracy while requiring less data for training. For example, advanced Transformers were employed for 3D medical image segmentation[21–23]. Hatamizadeh et al.[21] outlined a U-Net-based 3D medical image segmentation strategy that utilizes transformers as an encoder to learn the sequential representation of the inputs, while also adhering to the U-shaped network design. Hatamizadeh et al.[22] introduced a novel segmentation model (Swin-Unetr), redefining the task of semantic segmentation of the brain tumors as a sequence-to-sequence prediction challenge. Chen et al.[23] developed Transunet, which synergizes the strengths of Transformers and U-Net. It canachieve superior performance across various medical applications, albeit with limited generalization capabilities.

Alternative methods for medical image segmentation uses diffusion probabilistic models. For example, Wu et al.[24] introduced MedSegDiff, which eliminates high-frequency noise using an eigenfrequency analyzer and employs a dynamic conditional coding technique to establish adaptable conditions for each sampling step.

Medical image segmentation based on basic models

The widely used neural networks include CNNs, FCNs, U-Net, and Transformers. They are typically designed with particular objectives in mind. This often leads to limitations in their generalization ability. Meta introduced a generalized large model, the SAM, to achieve straightforward segmentation across various objects in images and videos. Following this, researchers have explored the application of SAM in medical image segmentation. For example, Mazurowski et al.[10] assessed three variations of the base SAM across 19 medical imaging datasets. The model excels at segmenting targets with distinct boundaries. However, in cases of indistinct boundaries, the box prompt method notably surpasses the point prompt approach. The performance based on a single prompt can greatly differ across various segmentation tasks.

Shi et al.[11] assessed SAM, finding it more effective on endoscopic and dermatoscopic imagery than on other types of medical images. However, SAM struggled with zero-sample segmentation in images featuring continuous branching structures, like blood vessels. Wald et al. reviewed the efficacy of SAM's larger models with both point and box prompts[12]. It found that SAM adapts effectively to CT imagery, yet it doesn't reach the pinnacle of segmentation performance. Wu et al. (2023) introduced MedSAM, tailored for extensive medical imaging datasets featuring numerous paired images and masks[9]. Experiments using the original SAM without fine-tuning confirmed SAM's effectiveness in datasets with well-defined targets and boundaries[9–12]. However, SAM faces challenges in most medical datasets characterized by fuzzy boundaries and small targets.

Meanwhile, researchers have concentrated on optimizing SAM for medical imaging. For example, Wu et al.[13] introduced the medical SAM adapter, a method that incorporates specialized medical knowledge into the segmentation model via a straightforward adaptation approach. It inserts the Adapter model at a specific location in the SAM network architecture to enable fine-tuning, achieving performance comparable to state-of-the-art methods. Zhang et al.[14] proposed SAMed, a versatile approach for medical image segmentation. This network applies a low-rank (LoRA) based fine-tuning strategy to the SAM image encoder, performing fine-tuning on labeled medical image segmentation datasets along with a prompt encoder and a mask decoder. According to the above[13,14], fine-tuning SAM improves its performance to a level comparable to current state-of-the-art segmentation methods. However, it does not address the segmentation speed. This limitation makes it challenging to quickly segment large numbers of images generated by imaging modalities such as CT. Therefore, researchers have focused on making SAM lighter without compromising its performance. For example, Zhao et al.[25] replaced the image encoder in the original SAM with yolov8-seg for training, using 2% of the original dataset for training and achieving a 50-fold speedup in SAM's performance. Zhang et al.[26] utilized a distillation method to replace the heavy image encoder with a lightweight image encoder, making it suitable for mobile applications. Xiong et al.[27] constructed EfficientSAM by leveraging a lightweight image encoder and a masked decoder, both pretrained on masked images. Zhao, Zhang, and Xiong provided ideas for a lightweight SAM, but no attempt has been made on medical image segmentation.

Efficient medical-images-aimed segment anything model

In this section, the EMedSAM will be presented in detail. The structure of EMedSAM is shown in Fig. 1, which consists of three main parts: a lightweight image encoder, a prompt encoder, and a mask decoder. First, medical images are preprocessed to enhance their quality for the encoding stage, which will be discussed in Sect. "Image preprocessing". Subsequently, the processed images are input into a lightweight image encoder integrating adapter technology, which effectively extracts key features from the images; this process is further elaborated in Sects. "Lightweight image encoder DD-TinyViT" and "Fine-tuning med-adapter", respectively. Once feature extraction is complete, the image feature embeddings are transmitted to the mask decoder, and similarly, embeddings from prompts are sent by the prompt encoder to the mask decoder. The mask decoder then integrates these embeddings and outputs the final segmented image. The overall training strategy of the model, especially how prompts are utilized to optimize the training process, will be detailed in Sect. "Adapter training".

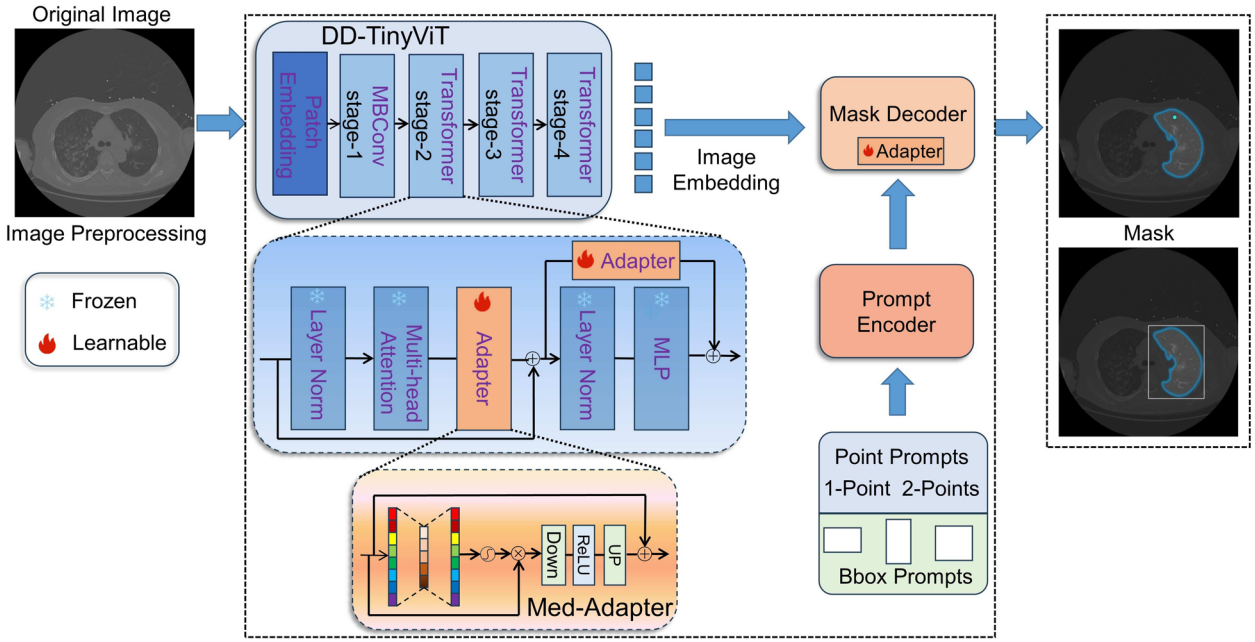


Figure 1. Structure of the EMedSAM.

Image preprocessing

Preprocessing of 2D medical imaging data

For data consistency and quality assurance, the dataset is preprocessed and used for subsequent analysis. In this paper, the first step of preprocessing is to apply a Gaussian filter for noise reduction, effectively separating valuable information from background noise. Then, the pixel values are scaled to a standard range through normalization to bolster the model's ability. To further enhance image quality, histogram equalization is used to augment contrast. To accommodate the model input, the size of all images is unified and stored in PNG format to ensure data consistency. Considering the limitation of the dataset size, enhancement operations such as rotation, scaling, panning, and flipping are applied to augment the dataset diversity.

Preprocessing of 3D medical imaging data

Preprocessing of the 3D medical image dataset begins with adjusting voxel values to the range of 0–255 through voxel normalization, ensuring uniformity for further analysis. Subsequently, histogram equalization is utilized to amplify the contrast and definition of voxel data. Then, the voxel size is modified or resampled to lessen computational demands while preserving consistency in the data. Finally, 2D slices are extracted from the 3D voxel data to facilitate analysis using 2D methods.

Lightweight image encoder DD-TinyViT

Image encoder structure

Image encoder is an essential part of SAM. The original SAM uses ViT-H as the backbone network of its image encoder and the number of parameters in ViT-H is up to 632M. This large parameter count requires significant computational cost when performing model training and inference, which escalates the hardware requirements and extends the processing time.

To address this issue, DD-TinyViT follows the TinyViT model framework, a lightweight image encoder with only 21M parameters, as an alternative to the original heavy encoder. Similar to Swin-Transformer, TinyViT is divided into five parts. The first part is an image embedding block using two convolutions with a kernel of 3 and a step size of 2. The second part is a novel bottleneck convolution block named MBConv, and the remaining part uses a sequence model Transformer based on the attentional mechanism. Downsampling operations are performed in each of the second to fifth parts. Each layer is connected using residuals, and the activation function is *GatedGeLU*:

$$GatedGeLU = GELU(x) \cdot \sigma(GELU(x)) \tag{1}$$

where x represents the input image, and σ is the Sigmoid function. The expressions of $\sigma(x)$ and $GELU(x)$ are as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

$$GELU(x) = 0.5x\left(1 + tanh\left(\sqrt{\frac{2}{\pi}}\left(x + 0.447x^3\right)\right)\right) \tag{3}$$

Lightweight handling

Direct training of smaller models on large datasets typically results in subpar performance. This paper utilizes the decoupled distillation technique to develop a lightweight image encoder, employing the pre-trained larger model (ViT-H) as the teacher model and the more compact model (DD-TinyViT) as the student model. The decoupled training procedure is shown in Fig. 2. Building upon Zhang's research[26], we transitioned the application domain from mobile applications to the medical field. We utilized a lightweight image encoder that outperforms MobileSAM, resulting in DD-TinyViT through decoupled distillation. Although this led to a slight increase in parameters, we implemented appropriate measures to address this issue. Unlike Zhao's FastSAM[25], which replaces ViT with YOLOv8 for speed advantages, we retained the original SAM model's ViT functionality. While YOLOv8 excels in processing speed, it falls short in comprehensive image content understanding. Preserving the ViT can ensure the model's depth and accuracy in image comprehension remain uncompromised.

During the distillation phase, the teacher model initially performs inferences on the data, with its output feature map being preserved. To reduce training duration, the SA-1B dataset undergoes preprocessing, and the feature map which is outputted by SAM is stored locally, eliminating the need for the SAM model's image encoder during training inference. Merely 1% of the SA-1B dataset is allocated for training, with 0.1% designated for validation. Following this, the student model is trained to minimize the discrepancy between its image embedding outputs and those of the teacher model, specifically the image embedding output $O_{ViT-H}(x_i)$, employing the Mean Square Error (MSE) as the loss function:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(O_V(x_i) - O_{DD-TV}(x_i)\right)^2 \tag{4}$$

where x_i represents the ith input image, N represents the total sample size, $O_V(x_i)$ represents the image embedding output from the ViT-H image encoder, and $O_{DD-TV}(x_i)$ represents the image embedding output from the DD-TinyViT image encoder.

To enhance the robustness of the student model and mitigate overfitting, L_2 regularization is used. incorporates a penalty term into the loss function, proportional to the square of the coefficient magnitudes. This penalty term promotes smaller weight values, which aids in reducing the model's tendency to overfit the training data:

$$L_{reg} = \lambda \sum_{\varpi \epsilon W} \varpi^2 \tag{5}$$

where ϖ represents the individual weights of the DD-TinyViT model, W represents the set of all weights within the DD-TinyViT model, and λ represents the regularization coefficient.

The overall loss function is:

$$L_{total} = L_{MSE} + L_{reg} \tag{6}$$

In contrast to more intricate loss functions like the combination of Focal and Dice losses, this method is straightforward and more amenable to optimization. The stochastic gradient descent (SGD) algorithm is employed to minimizing the total loss and facilitating the training of DD-TinyViT, with the update rule as follows:

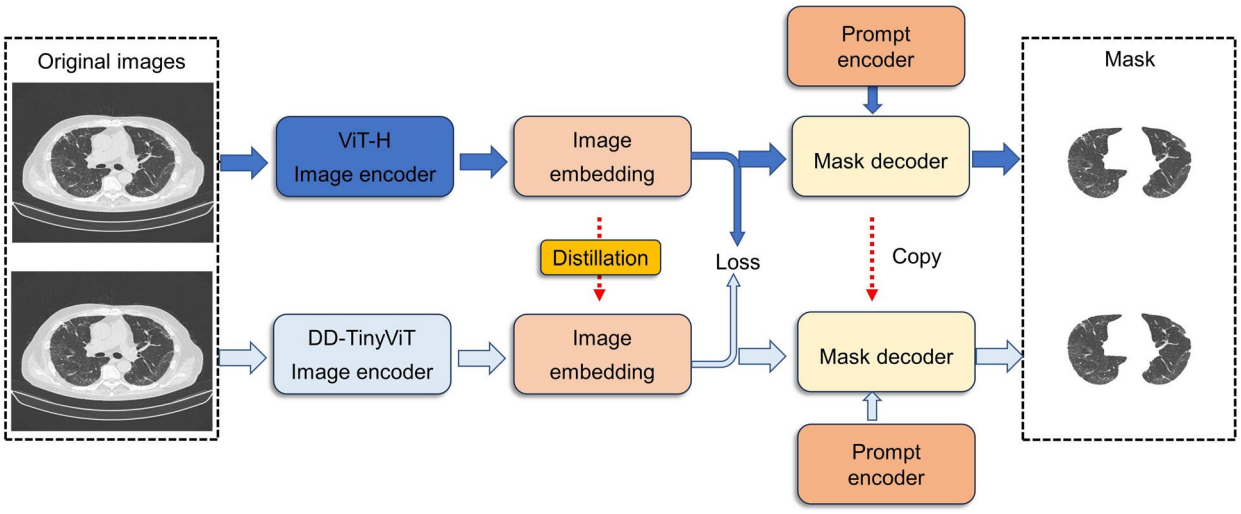


Figure 2. Decoupled distillation structure diagram.

$$\varpi_{t+1} = \varpi_t - \eta \nabla L_{total}(\varpi_t) \tag{7}$$

where ϖ_{t+1} and ϖ_t represents the updated and current weights of the model, respectively. η represents the learning rate, which determines the step size at each iteration in the quest to minimize the loss function. $\nabla L_{total}(\varpi_t)$ represents the gradient of the total loss with respect to the weights ϖ.

This training process is iterated as follows: if the total loss L_{total} exceeds a predefined threshold, the training of the DD-TinyViT image encoder continues; otherwise, the training concludes. This threshold is set based on DD-TinyViT achieving more than 90% of the original model ViT-H's performance. Subsequently, DD-TinyViT is evaluated to verify its performance equivalence with ViT-H. Through these steps, an optimized lightweight image encoder is obtained and integrated with the frozen prompt encoder and mask decoder.

Fine-tuning med-adapter

Adapter structure

Due to the limited availability of medical imaging data in the SA-1B large-scale dataset, the lightweight image encoder faces challenges in medical image segmentation tasks. To efficiently adapt this encoder for medical image segmentation, this paper introduces a novel adapter technology-med-adapter. Its design objective is to integrate specialized knowledge of medical imaging by embedding med-adapter within the Transformer layers, thereby preserving the previously acquired insights from the SA-1B dataset without the need for retraining. As shown in Fig. 3, this adapter features a bottleneck architecture. It initially reduces the dimensionality to a lower level, traverses a nonlinear activation function layer, and subsequently expands back to the original dimension. Moreover, a residual connection is maintained between the adapter layer's input and output to guarantee uninterrupted information flow.

Adapter deployment

Considering that the configuration of med-adapter units can influence the overall efficacy of the model. Building on the Medical SAM Adapter design proposed by Wu[13], this study embarked on multiple experimental investigations to assess the impact of med-adapter placements within different sections of the model. For specific details, refer to the third paragraph of the "Ablation Study of EMedSAM" section. We determined the placement of the med-adapter from multiple sets of experiments, further validating Wu's findings.

In the encoder segment, for 2D medical imaging, this paper implements two adapters within each Transformer block of DD-TinyViT, as shown in Fig. 3a. The first adapter is situated downstream of the multi-head attention module yet prior to its associated residual connection. This placement allows the adapter to refine and enhance the module's outputs without disrupting the main information flow, thereby improving the precision and efficiency of the information processing.The second adapter is positioned within the residual pathway of the MLP layer that ensues the multi-head attention component. By introducing an adapter at this location, the model leverages the computational power of the MLP layer to increase the nonlinear capabilities during the processing, thereby enhancing the model's ability to learn complex data patterns.

For 3D medical imaging, one approach involves using image preprocessing to convert the images into 2D slices, employing the adaptation method described in Fig. 3a. Another approach utilizes the configuration

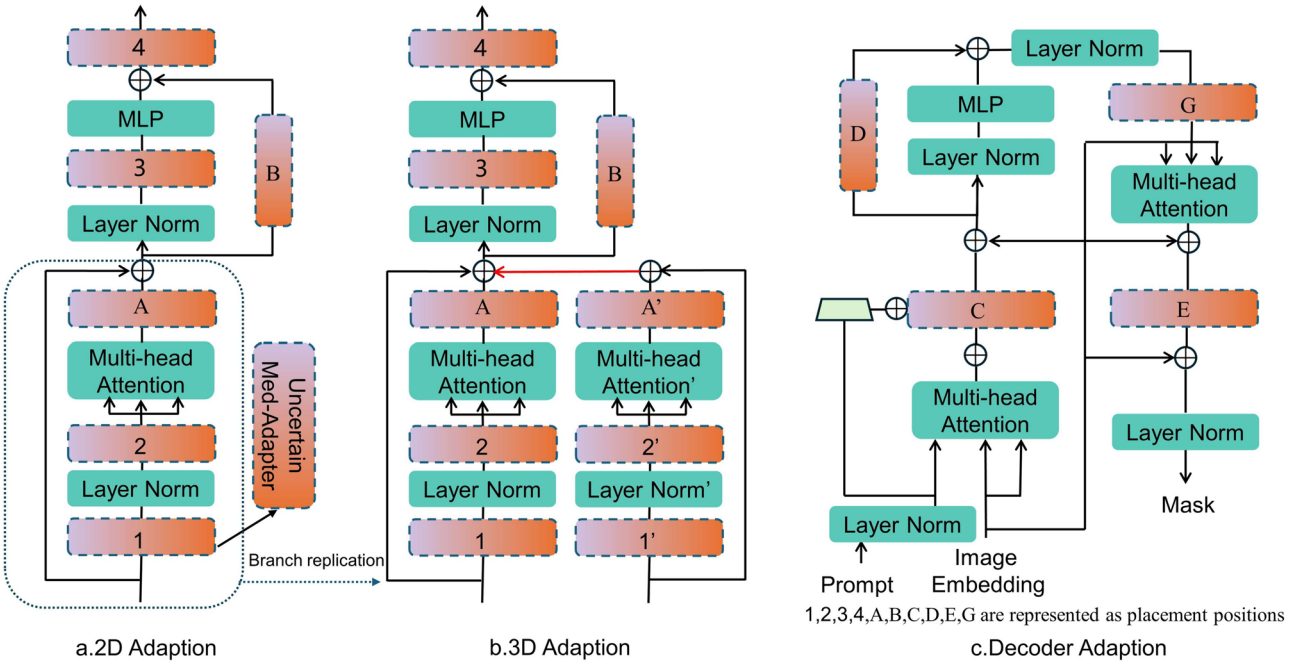


Figure 3. Deployment structure diagram of med-adapter.

depicted in Fig. 3b, where the placement of the adapters remains the same, but an additional branch is introduced at the bottom. Specifically, 3D medical imaging, due to their higher spatial dimensions, contain more information and contextual relationships. Consequently, more complex adapter structures are required to effectively extract and process these additional data dimensions. In contrast, 2D medical imaging can achieve efficient handling with relatively simpler adapter structures because of their lower data dimensionality. The original attention mechanism is divided into spatial and depth branches. For 3D samples with a depth of Z, the spatial branch learns spatial correlations through interactions of M×P, where M is the number of embeddings, and P is the length of embeddings. The depth branch learns depth correlations through Z×P interactions. Ultimately, the processed outputs from the depth branch are restored to their original form and combined with the results from the spatial branch.

In the decoder segment, three adapters are arranged for each Transformer block. The initial adapter is located ubsequent to the multi-head cross-attention module, which transforms prompt embeddings into image embeddings and adds a residual connection. This positioning optimizes the interplay of attention and residual pathways, enhancing prompt integration. An additional down-projection step, implemented before the ReLU activation function, condenses the prompt embeddings to enhance their processing. This setup enables the adapter to adeptly adjust parameters based on enriched prompt information, thereby increasing its adaptability across different modalities and tasks.

The decoder's second adapter, akin to its encoder counterpart, is adeptly placed to adjust the MLP-augmented embeddings, further refining the model's ability to handle detailed and complex data patterns. The third adapter is positioned at the residual junction subsequent to the cross-attention from image embeddings to prompt attention. Following these adapter connections, an additional residual connection and layer normalization are introduced to stabilize the output and ensure uniformity, crucial for high accuracy in complex segmentation tasks.

Adapter training

To achieve optimal performance, fine-tuned adapter parameters to medical imaging datasets are essential. During the training process, the other parameters of EMedSAM are kept frozen, and only the adapter parameters are updated. The essence of the fine-tuning strategy lies in optimizing the channel dimension and spatial dimension. We comprehensively optimized the model by dynamically adjusting the parameters across channel and spatial dimensions, in contrast to Wu's research, which was confined to optimizing the spatial and depth branches of the training process for the 3D medical adapter. For the channel dimension, initially, the resolution of the input feature map is reduced through global average pooling, averaging the information across the entire feature map to produce a global, more compact representation. This strategy diminishes the model's parameter count while preserving essential feature information:

$$F_{pool} = \frac{1}{N}\sum_{i=1}^{N} F_i \tag{8}$$

where F_{pool} represents the outcome post the global average pooling operation. N represents the total number of elements in the feature map, with each element F_i corresponding to a unique position within this flattened vector representation.

Subsequently, a linear layer is employed to refine the channel embeddings, which effectively narrows the spatial dimensions of the data and isolates the principal features:

$$F_{comp} = Linear\left(F_{pool}\right) \tag{9}$$

where F_{comp} represents the condensed output which is derived following the linear transformation and $Linear$ represents the linear transformation operation, encapsulating a fully connected layer. The layerencompasses a weight matrix and a bias term, executing a linear combination on the input.

Moreover, an additional linear layer is employed to revert the data back to its original dimensionality:

$$F_{rest} = Linear\left(F_{comp}\right) \tag{10}$$

where F_{rest} represents the restored compressed embedding channels.

Finally, weights for the channel dimension are obtained through the sigmoid function and multiplied with the input feature map, serving as the input for the next stage:

$$W_{chan} = Sigmoid(Linear(F_{rest})) \tag{11}$$

$$F_{next} = W_{chan} \odot F_{pool} \tag{12}$$

where W_{chan} represents the output achieved by executing a linear transformation on F_{rest} and integrating an activation function. Sigmoid is the chosen activation function. $\odot$ represents element-wise multiplication, and F_{next} is the output realized by applying element-wise multiplication to F_{pool}, with the weights being W_{chan}.

For the spatial dimension, a convolutional layer is employed to reduce the spatial resolution of the feature map to half of its original size, effectively downsampling the input features and eliciting new feature representations. This technique is instrumental in capturing the spatial structural information inherent in the input:

$$F_{downs} = Conv2D(F_{next}) \tag{13}$$

where F_{downs} represents the results from applying a 2D convolution operation to F_{next}.

Utilizing transposed convolution, the spatial resolution gets reinstated while preserving the input's channel count. This process expands the spatial dimensions of the input features, effectuating upsampling.

$$F_{upsa} = TransposeConv2D(F_{downs}) \tag{14}$$

where F_{upsa} is the result achieved by executing a transposed convolution operation on F_{downs}, where *TransposeConv2D* represents the transposed convolution operation.

Post-adaptation in each layer, a skip connection is introduced to integrate the two feature representations. This approach bolsters information flow, retains finer details, and facilitates the transfer of information along with the propagation of gradients.

$$F_{skip} = F_{next} + F_{upsa} \tag{15}$$

where F_{skip} is derived by executing element-wise addition between F_{next} and F_{upsa}, culminating in an output that aligns with the image encoder specifically adjusted for the current medical imaging data.

This paper employs standard loss functions, including cross-entropy and Dice loss, to refine the model. To avert overfitting and facilitate the adapter's acquisition of medical domain expertise, regularization methods like weight decay and Dropout are implemented. This approach ensures the model's focus on assimilating precise knowledge pertinent to the medical field during training. Through the integration of adapters and targeted fine-tuning, the lightweight SAM has realized noteworthy enhancements in performance across a variety of medical image segmentation tasks.

Overall training strategy

Inspired by SAM training, the bounding box prompts remain unchanged. As for the point prompts, our method employs a technique that amalgamates random sampling with iterative sampling. In the initial phases of model training, click prompts are commenced with random sampling, aiding the model in the early discernment of foreground and background regions within the image. Subsequently, to boost the model's accuracy and adaptability, an iterative sampling mechanism is incorporated. In this regimen, each click iteration is informed not solely by the model's extant predictions but also by inaccuracies identified in prior outputs, thus directing new clicks toward these discrepancies. This iterative sampling approach is crafted to mimic real-world scenarios where users incrementally refine model forecasts through ongoing interaction. Via persistent iteration, the model progressively enhances its comprehension of intricate image constituents and segmentation fidelity.

Through the hybrid training strategy that merges random sampling with iterative click sampling, the model can grasp the fundamental structure of the image initially, and continuously optimize and refine segmentation outcomes in subsequent training phases. Then it improves accuracy and user interaction experience.

Experimental results and analysis

In this section, we conduct a series of experiments to thoroughly assess the EMedSAM's capabilities in medical image segmentation tasks.

Dataset

In this study, we conduct experiments on four different medical image segmentation datasets. Our goal is to thoroughly showcase the accuracy, efficiency, and versatility of the EMedSAM. The model is deployed across a range of medical imaging modalities for intricate segmentation challenges. The datasets utilized are categorized into two principal groups: One group assesses the model's overall segmentation capabilities, while the other examines its adaptability across diverse medical imaging types.

Specifically, we choose the MICCAI FLARE 2022 challenge dataset for assessing abdominal multi-organ segmentation, featuring 2300 3D CT scans from over 20 institutions, including the liver, kidney, and spleen. Two datasets from the First Hospital of Zhejiang University School of Medicine and one public dataset are employed to evaluate the model's versatility in various imaging modalities. The first is a lung CT scan segmentation dataset comprising 1500 2D images. and the second is a mammography X-ray segmentation dataset with 1200 2D images. The third is the BraTs dataset which is a public repository for brain MRI tumor segmentation. Each dataset is meticulously annotated with the relevant segmentation masks.

Experiment details

In this study, we train the EMedSAM to ensure its optimal performance in medical image segmentation tasks. We employ meticulous training setups and experimental schemes to rigorously assess the EMedSAM across various medical imaging segmentation challenges and investigate the influence of interactive prompts on learning efficacy and accuracy. Model training leverages the DD-TinyViT framework, entailing precise adjustments to the SAM's encoder and incorporating med-adapter to enhance the assimilation of medical image-specific insights and details. The conventional SAM training setup is applied for 2D medical imaging tasks, while for 3D imaging tasks, we set the batch size to 8 to prevent memory overflow. To maintain experimental comparison fairness, for the comparative models, we did not simply fix the number of training epochs. Instead, we closely monitored each model's performance on the validation set, including the loss curves and accuracy. We selected the model that performed best on the validation set for evaluation on the test set, ensuring that the chosen model exhibited optimal performance. For EMedSAM, when using the med-adapter, a shorter training duration was chosen. Specifically, we conducted training for 50 epochs uniformly across all datasets. This decision was based on our observation that, for EMedSAM, 50 epochs were sufficient to achieve stable and excellent training

outcomes. Model parameter optimization is conducted using the Adam algorithm, setting the encoder adapter's hyperparameters to $\alpha = 0.5$ and $\beta = 0.1$.

During interactive training, three distinct prompt mechanisms are investigated to gauge their effect on model performance: (1) A single positive point prompt (1-point) to denote the foreground interest area; (2) Double positive point prompts (2-points) offering extra spatial localization info; and (3) Bounding box prompts (BBox) overlapping the target area to delineate the entire region of interest. All experiments are conducted within the PyTorch framework, utilizing two NVIDIA A100 GPUs for training and evaluation. Test dice scores in the experimental outcomes are derived from the best-performing models during validation.

Evaluation indicators

To conduct an in-depth assessment of the EMedSAM's effectiveness in medical image segmentation, we utilize key metrics such as the dice similarity coefficient (DSC), the 95% Hausdorff distance (HD95), and the intersection over union (IoU). These criteria provide a multi-faceted evaluation of the concordance and precision between the predicted outcomes of the model and the ground truth annotations, establishing a solid foundation for a detailed analysis of the model's performance.

The DSC measures the extent of agreement between the predicted segmentation and the ground truth, with its values ranging from 0 (no overlap) to 1 (perfect agreement). The DSC is computed as follows:

$$DSC(X, Y) = \frac{2\|X \cap Y\|}{|X| + |Y|} \tag{16}$$

where, X represents the predicted segmentation area, Y represents the actual segmentation area, and $X \cap Y$ represents the overlap between them.

HD95 is derived from the 95th percentile of the distances between the perimeters of two point sets, acting as a reliable indicator for gauging boundary segmentation precision. The computation of HD95 is as follows:

$$HD95(A, B) = 95th\ percentile\ of\ \left\{ dist(a, B) \cup dist(b, A) \mid a \in A, b \in B \right\} \tag{17}$$

where, A and B represent the boundary point sets of the prediction and actual annotations, respectively, and $\mathrm{dist}(a, B)$ calculates the shortest distance from point a to every point in set B.

The IoU metric measures the similarity between the predicted and true regions. IoU values span from 0 to 1, where the value 1 signifies an exact correspondence. IoU is determined by the following equation:

$$IoU(X, Y) = \frac{\|X \cap Y\|}{\|X \cup Y\|} \tag{18}$$

where, $X \cap Y$ represents the intersection between the predicted region and the actual region, whereas $X \cup Y$ represents their union.

Evaluation of EMedSAM

Quantitative comparison

To fully assess the comprehensive performance of the EMedSAM in the field of medical image segmentation, we offer a detailed comparison with the current state-of-the-art (SOTA) segmentation methods on the FLARE 2022 multi-organ segmentation dataset. Quantitative analysis outcomes are encapsulated in Table 1, showcasing EMedSAM's performance relative to various esteemed medical image segmentation algorithms, such as nnU-Net[19], Transunet[23], UNetr[21], Swin-Unetr[22], Medsegdiff[24], inclusive of the standard SAM and the fully optimized MedSAM model. Herein, the Dice coefficient serves as the primary metric for evaluating each model's performance in the multi-organ segmentation process.

Table data reveals that EMedSAM achieves notable performance over conventional SAM models with just a single-point prompt. Especially, EMedSAM attains SOTA-level segmentation outcomes for 13 principal organs in the FLARE 2022 dataset, surpassing all other benchmarked methods collectively. When we introduce more detailed prompt inputs, such as double-dot and box prompts, it further boosts EMedSAM's performance, achieving an average Dice coefficient up to 88.9%. It is a 2.7 percentage point higher than the SOTA model, Swin-Unetr. Therefore, EMedSAM delivers superior results with a lighter architecture, and is efficiency and practicality.

Then, we analyze the reactions of interactive segmentation models, such as SAM, MedSAM, and EMedSAM, to varied prompt types. It reveals that double-dot prompts generally excel over single-dot prompts. This implies that with an increase in valid prompts, the model's capacity to pinpoint the segmentation target enhances correspondingly, thereby refining segmentation outcomes. Regarding box prompts, it's observed that they don't always surpass point prompts in SAM segmentation, possibly varying with organ characteristics; nonetheless, box prompts exceed single-point prompts in the majority of instances. In the MedSAM, box prompts are not always superior or on par with dot prompts. Conversely, in EMedSAM, segmentation performance steadily improves with enriched prompt information, and box prompts almost consistently outperform point prompts. Furthermore, accurate bounding box labeling is pivotal for performance improvements as it explicitly delineates the segmentation target's boundaries. Therefore, all interactive models exhibit consistent, incremental performance enhancements with the addition of effective prompt data across various prompt scenarios, and box prompts typically surpass single-point prompts. This further confirms that the models are high sensitivity and adaptability to prompt inputs. In evaluating the benchmark performance of SAM within medical image segmentation tasks, no matter what prompt is usedi, SAM generally underperform in zero-sample scenarios compared to fully trained MedSAM. Although this comparison might appear inequitable due to the different training methodologies, it

Model	Param(M)	Liver	Kidney-R	Spleen	Pancreas	Aorta	Ivc	Rag
nnU-Net[19]	16	0.942	0.902	0.938	0.691	0.882	0.787	0.625
Transunet[23]	37	0.963	0.925	0.948	0.778	0.923	0.828	0.64
UNetr[21]	104	0.97	0.926	0.965	0.763	0.887	0.851	0.745
Swim-Unetr[22]	138	0.973	0.938	0.97	0.79	0.895	0.855	0.76
Medsegdiff[24]	32	0.951	0.935	0.949	0.779	0.851	0.829	0.719
SAM 1 point[8]	636	0.456	0.652	0.507	0.523	0.62	0.419	0.358
MedSAM 1 point[9]	636	0.893	0.821	0.753	0.698	0.87	0.749	0.71
Ours 1 point	21	0.969	0.932	0.971	0.763	0.91	0.87	0.801
SAM 2 points[8]	636	0.492	0.702	0.613	0.561	0.647	0.478	0.392
MedSAM 2points[9]	636	0.91	0.828	0.76	0.761	0.876	0.753	0.715
Ours 2 points	21	**0.978**	0.935	0.977	0.774	0.916	0.877	0.814
SAM BBox[8]	636	0.473	0.638	0.501	0.579	0.623	0.55	0.35
MedSAM BBox[9]	636	0.902	0.84	0.761	0.754	0.891	0.748	0.709
Ours BBox	21	0.976	**0.942**	**0.981**	**0.79**	**0.934**	**0.89**	**0.819**
Model	**Param(M)**	**Lag**	**Gall**	**Esophagus**	**Stomach**	**Duodenum**	**Kidney-L**	**Avg**
nnU-Net[19]	16	0.619	0.712	0.73	0.827	0.725	0.91	0.792
Transunet[23]	37	0.635	0.659	0.754	0.894	0.787	0.93	0.82
UNetr[21]	104	0.739	0.753	0.77	0.918	0.783	0.939	0.847
Swim-Unetr[22]	138	0.767	0.798	0.778	0.925	0.809	0.945	0.862
Medsegdiff[24]	32	0.726	0.733	0.769	0.926	0.795	0.926	0.838
SAM 1 point[8]	636	0.352	0.535	0.59	0.56	0.55	0.788	0.532
MedSAM 1 point[9]	636	0.705	0.77	0.725	0.851	0.768	0.882	0.784
Ours 1 point	21	0.798	0.818	0.809	0.928	0.81	0.957	0.872
SAM 2 points[8]	636	0.385	0.608	0.601	0.668	0.621	0.81	0.583
MedSAM 2points[9]	636	0.712	0.779	0.73	0.856	0.774	0.887	0.795
Ours 2 points	21	0.81	0.824	0.817	0.933	0.812	0.96	0.879
SAM BBox[8]	636	0.35	0.616	0.6	0.558	0.659	0.735	0.556
MedSAM BBox[9]	636	0.705	0.775	0.747	0.858	0.763	0.879	0.795
Ours BBox	21	**0.815**	**0.838**	**0.82**	**0.939**	**0.847**	**0.968**	**0.889**

Table 1. Model performance test results based on FLARE 2022 dataset. Significant values are in bold.

uncovers SAM's constraints in generalizing to medical image segmentation tasks under zero-sample conditions. This observation aligns with prior studies, highlighting the necessity for specialized techniques to refine the SAM to meet the complex requirements of medical image segmentation.

In subsequent analyses, we conducted a comprehensive comparison of the EMedSAM across diverse imaging modalities in medical image segmentation to thoroughly evaluate its performance. Outcomes from this comparative study are elaborately presented in Tables 2, 3 and 4, employing pivotal performance indicators like Dice

Model	IoU	DSC	HD
Unetr[21]	80.3	87.6	12.63
nn-UNet[19]	80.8	88.9	11.41
Swin-Unetr[22]	81.2	88.1	11.52
Transunet[23]	79.3	87.2	13.54
Medsegdiff[24]	78.2	85.1	14.25
SAM 1 point[8]	51.5	65.8	33.58
MedSAM 1 point[9]	75.6	82.4	16.32
Ours 1 point	81.3	88.2	11.42
SAM 2 points[8]	63.4	72.3	29.54
MedSAM 2points[9]	76.1	83.2	15.56
Ours 2 points	82.1	89.3	11.35
SAM BBox[8]	63.6	75.6	28.32
MesSAM BBox[9]	76.5	84.1	15.10
Ours BBox	**82.4**	**90.1**	**10.21**

Table 2. Model performance test results based on lung CT dataset. Significant values are in bold.

Model	IoU	DSC	HD
Unetr[21]	79.8	87.4	11.48
nn-UNet[19]	80.3	88.7	11.39
Swin-Unetr[22]	81.5	87.9	11.44
Transunet[23]	78.6	85.6	14.10
Medsegdiff[24]	79.6	87.1	13.35
SAM 1 point[8]	52.3	66.3	32.63
MedSAM 1 point[9]	76.3	81.2	15.42
Ours 1 point	81.6	87.6	12.46
SAM 2points[8]	62.8	71.6	28.24
MedSAM 2points[9]	76.9	82.8	15.50
Ours 2 points	82.5	88.2	12.10
SAM BBox[8]	64.2	74.5	24.31
MesSAM BBox[9]	77.8	84.3	16.10
Ours BBox	**83.1**	**89.1**	**11.23**

Table 3. Model performance test results based on breast mammography X-ray mass segmentation dataset. Significant values are in bold.

Model	IoU	DSC	HD
Unetr[21]	80.4	87.2	12.83
nn-UNet[19]	80.6	88.5	11.20
Swin-Unetr[22]	81.6	88.6	11.32
Transunet[23]	78.9	86.5	13.71
Medsegdiff[24]	77.3	85.4	14.29
SAM 1 point[8]	47.2	63.5	31.98
MedSAM 1 point[9]	74.5	81.6	15.56
Ours 1 point	81.0	88.4	11.15
SAM 2points[8]	64.3	70.5	29.54
MedSAM 2points[9]	74.9	82.1	15.03
Ours 2 points	82.3	89.3	10.11
SAM BBox[8]	62.5	75.6	26.37
MesSAM BBox[9]	76.1	83.4	14.90
Ours BBox	**82.6**	**89.5**	**9.68**

Table 4. Model performance test results based on brain MRI tumor segmentation dataset. Significant values are in bold.

coefficients, IoU, and HD95 for an all-encompassing assessment. Analytical findings reveal that traditional SAM grapple with medical images featuring ambiguous boundaries, especially in tasks like lung segmentation, breast mass segmentation, and brain tumor segmentation. Conversely, while the fully trained MedSAM exhibits good performance in numerous segmentation tasks, it shows restricted performance on 3D images, with its capability in brain tumor segmentation being notably inadequate. This shortfall underscores the pressing demand for algorithms proficient in handling complex 3D medical images. EMedSAM showcases superior performance across all assessed segmentation tasks, effectively attaining proficient generalization across an extensive array of medical imaging modalities and segmentation tasks. Especially, in the renowned BraTs benchmark for brain tumor segmentation, leveraging its optimization for 3D imaging, EMedSAM secures a 0.9% improvement in Dice coefficient relative to the SOTA model, Swin-Unetr, alongside a significant enhancement of 1.64% in the HD95 measure. These outcomes not only underscore EMedSAM's excellent accuracy, robustness, and adaptability but also its progress and applicability in addressing medical image segmentation tasks, offering robust technical backing for forthcoming research and clinical efforts in medical image segmentation.

A detailed evaluation of processing speeds among EMedSAM, SAM, and fully trained MedSAM is conducted, with outcomes shows in Table 5. Additionally, comparisons were made with FastSAM, MobileSAM, and EfficientSAM. Given that MedSAM is predicated on the SAM architecture and refined with an extensive corpus of medical imagery, its processing velocity is similar to that of the original SAM. Nevertheless, the integration of a lightweight image encoder in EMedSAM results in a marked decrement in parameter count, considerably elevating the model's functional efficacy. Compared to purely lightweight efficiency models, our model demonstrates a significant speed improvement over FastSAM while maintaining parity with MobileSAM and EfficientSAM.

Model	SAM[8]	MedSAM[9]	FastSAM[25]	MobileSAM[26]	EfficientSAM[27]	Ours
Param(M)	636	636	68	10	9.8	21
Speed(ms)	456	456	40	10	8	12

Table 5. Model parameters and running speed.

Although the time difference is only a matter of milliseconds, it is noteworthy that EMedSAM is not solely an efficiency-first model.

Particularly, EMedSAM's theoretical processing speed has markedly increased, dropping from an initial 456 ms/image to a mere 12 ms/image, a substantial acceleration largely due to the lightweight image encoder's efficiency. While the actual execution speed varies with the employed GPU's capability, real-world assessments show EMedSAM to be approximately 60% faster than both SAM and MedSAM. This considerable speedup notably boosts efficiency without sacrificing on performance, particularly in situations demanding the processing of extensive medical imagery. These findings not only proves the EMedSAM's elevated efficiency in medical image segmentation but also illustrate how dual enhancements in processing speed and performance, motivated by technological advancements, create new opportunities for future medical image analysis. Particularly in medical contexts with stringent real-time processing demands, this benefit is poised to foster the broad adoption of EMedSAM in healthcare settings.

Qualitative comparison

We performed a qualitative analysis by visually comparing the segmentation results of EMedSAM with the SAM method . As shown in Fig. 4, EMedSAM's segmentation masks are notably smoother and more accurate compared to those produced by other methods.With the enrichment of prompt information, the segmentation performance of both SAM and EMedSAM has significantly improved. However, in dealing with small targets characterized by blurry segmentation boundaries, SAM exhibits inferior segmentation performance, whereas EMedSAM exhibits excellent segmentation capabilities. This marked enhancement is caused by two primary aspects:

Firstly, the large-scale SAM boasts superior feature extraction capabilities. It can effectively capture subtle features and complex patterns in images. This capability is crucial for medical image segmentation, given the extensive detailed content and subtle structural variations typical in medical images. Utilizing this strength, EMedSAM produces more accurate and detailed segmentation outcomes.

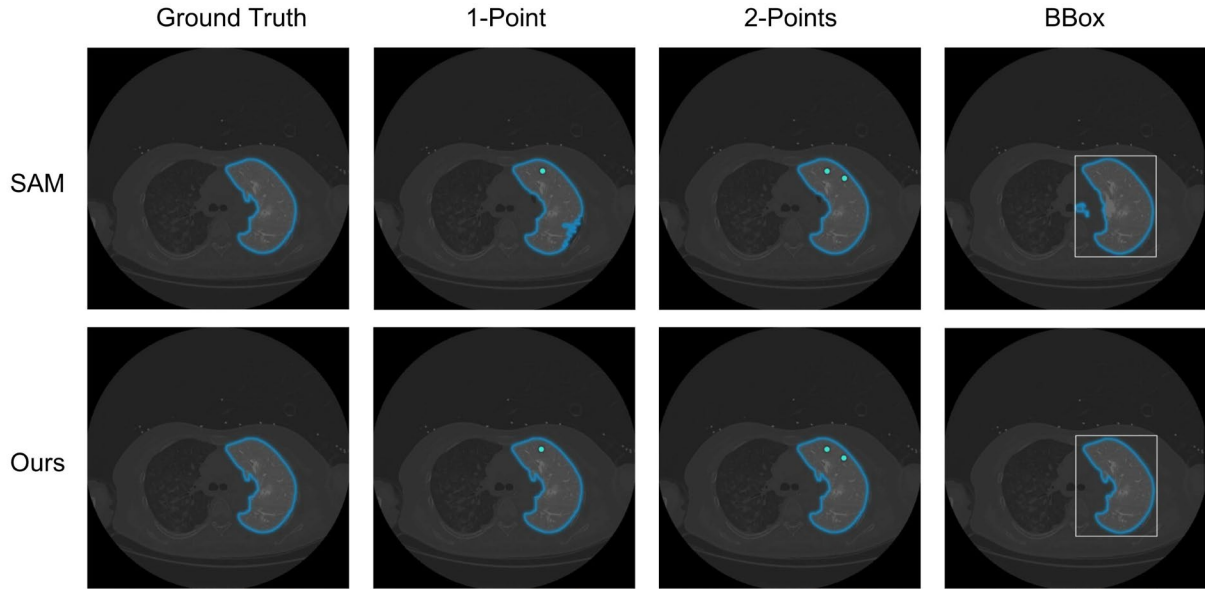
Secondly, the fine-tuning approach employed by EMedSAM is also instrumental. A well-conceived fine-tuning protocol, customized for features of medical imaging, enables the model to more adeptly conform to the distinct attributes of medical images and segmentation task demands. This fine-tuning guarantees the model's efficiency and accurate in feature extraction from medical images, thereby elevating the segmentation results' quality.

Ablation study of EMedSAM

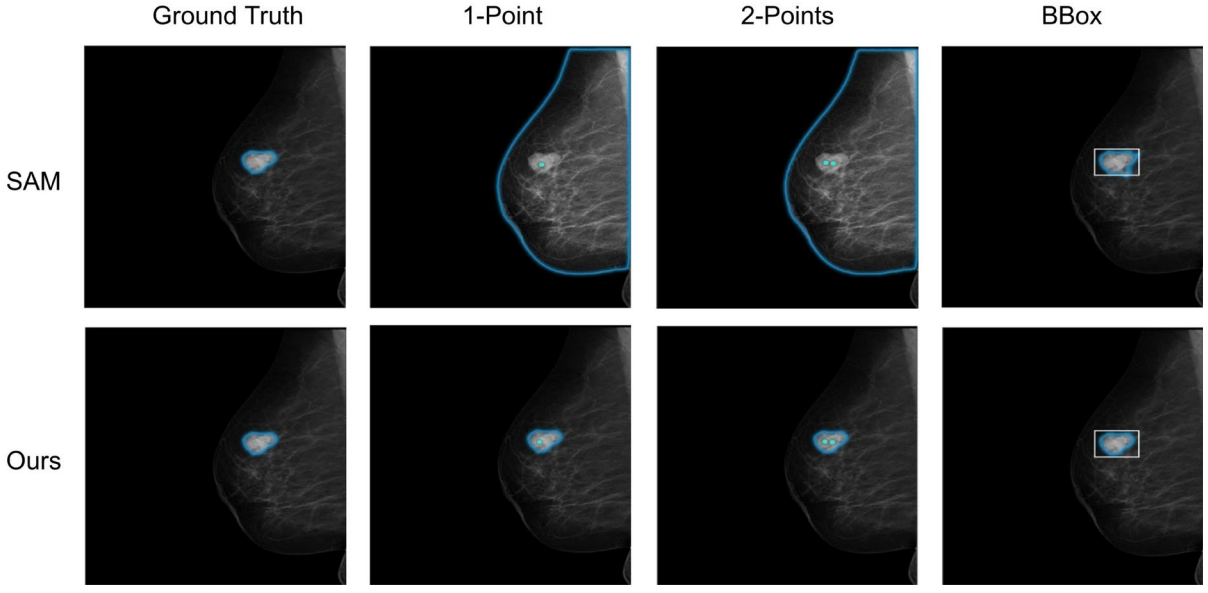
Through ablation studies, we comprehensively evaluate the specific impact of essential enhancements incorporated into the SAM. Such improvements encompass novel image encoders and adapters. Through the strategic removal or replacement of core components within the EMedSAM, including the integration of cutting-edge lightweight image encoders, such as DD-TinyViT, and the model-specific parametric adapter (med-adapter), we demonstrate these novelties' roles in enhancing the model's performance. Their influences are evaluated both individually and collaboratively.

We first use the original SAM as a baseline to evaluate the segmentation performance. Subsequently, we undertake an image encoder ablation study. It substitutes SAM's original image encoder with the proposed lightweight encoder (DD-TinyViT) and donot integrates an adapter, maintaining the model's remaining parts unchanged. This step evaluates the effect of the new image encoder in improving segmentation accuracy. Subsequently, we perform a parameter adapter ablation, retaining the original image encoder while solely adding the parameter adapter (med-adapter). This evaluates the adapter's efficacy in individually conforming to medical images' characteristics. Finally, we evaluate the entire model by integrating both the new image encoder and the parameter adapter into the SAM simultaneously. This forms the fully-trained EMedSAM, showcasing the collective impact of all enhancements. Ablation study findings presented in Table 6 reveal that substituting SAM's original image encoder with the lightweight DD-TinyViT doesn't markedly diminish segmentation accuracy relative to the baseline. This validates DD-TinyViT's proficiency in feature extraction and processing within medical imagery, signifying that adopting a lightweight image encoder enhances the model's velocity while preserving its original effectiveness. Furthermore, incorporating med-adapter improves the model's adaptability to medical images' characteristics, elevating segmentation's accuracy and consistency. This highlights the crucial role of the adapter in adjusting and optimizing the model for specific medical image segmentation tasks. Most notably, the most significant improvement is observed when integrating both the lightweight image encoder and the parameter adapter improvements into the EMedSAM. This demonstrates the synergistic effect of the individual component improvements and underscores their importance in enhancing the efficiency and accuracy of medical image segmentation tasks.

Additionally, this study conducted multiple experimental investigations to assess the impact of med-adapter placements within various sections of the model. Figure 3a illustrates the experimental setup for the encoder, where, to maintain experimental integrity, adjustments were confined solely to the positions and quantities of the encoder's adapters, while the decoder remained unchanged. As shown in Table 7, the adapter positions, denoted



(a) Lung CT segmentation



(b) Mammography X-ray mass segmentation

Figure 4. Visual comparison of EMedSAM and SAM in medical image segmentation.

Model	SAM	SAM+TinyViT	SAM+Adapter	EMedSAM
Image encoder	ViT-H	DD-TinyViT	ViT-H	DD-TinyViT
Parameter adapter	No adapter	No adapter	Med-adapter	Med-adapter
DSC (%)	58.3	58.1	87.5	88.9
Relative velocity (times)	1.0	1.6	1.0	1.6

Table 6. Ablation study of DD-TinyViT and med-adapter on the FLARE 2022 dataset.

as 1, 2, 3, 4, A, and B, were systematically explored through various combinations. Our findings revealed a marked performance decline in the absence of any adapters, and configurations with a single adapter were significantly less effective than those with dual adapters. The most effective performance was observed in the combination of positions A and B. Notably, the addition of more than two adapters did not yield further performance benefits but led to a considerable increase in computational resource consumption. Continuing with the experimental protocol, Fig. 3c illustrates the experimental setup for the decoder, utilizing the optimal adapter configuration

Model variant	Image encoder	Adapter configuration	DSC(%)
Baseline AB	DD-TinyViT	A + B	88.90
Single A	DD-TinyViT	A	81.21
Single B	DD-TinyViT	B	82.53
Variant 1A	DD-TinyViT	A + 1	84.50
Variant 2A	DD-TinyViT	A + 2	84.83
Variant 3A	DD-TinyViT	A + 3	85.52
Variant 4A	DD-TinyViT	A + 4	85.31
Variant 1B	DD-TinyViT	B + 1	86.00
Variant 2B	DD-TinyViT	B + 2	86.32
Variant 3B	DD-TinyViT	B + 3	85.80
Variant 4B	DD-TinyViT	B + 4	86.50
Variant 1AB	DD-TinyViT	A + B + 1	88.95
Variant 2AB	DD-TinyViT	A + B + 2	88.83
Variant No	DD-TinyViT	No	60.32

Table 7. Ablation study on the placement and quantity of med-adapter in the encoder on the FLARE 2022 dataset.

from the encoder for subsequent investigations. As shown in Table 8, we evaluated the impact of adapter placements at positions C, D, E, and G through additional combination experiments. The results confirmed that the absence of adapters resulted in the lowest performance, highlighting the critical role of adapters in enhancing system efficiency. While the introduction of a single adapter led to some performance improvements, the most significant enhancements were seen with combinations of three adapters, particularly at positions C, D, and E. However, adding more adapters beyond this configuration did not lead to substantial gains, suggesting diminishing returns and questioning the cost-effectiveness of further adapter integration.

Overall, the above experimental results show that by introducing the lightweight DD-TinyViT image encoder, the processing speed of the model is substantially improved while maintaining high-quality segmentation. The addition of med-adapter further optimizes the segmentation performance of the model, making it more adaptable to the characteristics of medical images. These improvements allow the EMedSAM to exhibit exceptional performance in medical image segmentation tasks. This model combines speed with accuracy, offering robust backing for prospective applications.

Discussion and conclusion

This study explores the feasibility of customizing large-scale models for medical image segmentation tasks. Through experiments, we show that large-scale fully-developed medical image segmentation models, when fine-tuned thoughtfully, can surpass the performance of conventional. We propose EMedSAM, a novel and fast segmentation model specifically designed for medical images. It incorporates the DD-TinyViT image encoder with the parameter adaptation capabilities of med-adapter. Utilizing a lightweight image encoder significantly

Model variant	Image encoder	Adapter configuration	DSC(%)
Baseline CDE	DD-TinyViT	C + D + E	88.90
Variant CDG	DD-TinyViT	C + D + G	87.58
Variant DEG	DD-TinyViT	D + E + G	86.83
Variant CDEG	DD-TinyViT	C + D + E + G	88.90
Single C	DD-TinyViT	C	81.20
Single D	DD-TinyViT	D	81.63
Single E	DD-TinyViT	E	80.81
Single G	DD-TinyViT	G	80.35
Variant CD	DD-TinyViT	C + D	84.10
Variant CE	DD-TinyViT	C + E	84.32
Variant CG	DD-TinyViT	C + G	83.42
Variant DE	DD-TinyViT	D + E	85.00
Variant DG	DD-TinyViT	D + G	84.52
Variant EG	DD-TinyViT	E + G	83.65
Variant No	DD-TinyViT	No	79.60

Table 8. Ablation study on the placement and quantity of med-adapter in the decoder on the FLARE 2022 dataset.

boosts the model's operational efficiency. Simultaneously, med-adapter's adaptive parameter learning enables the model to more accurately identify and segment target structures in medical images. Our strategy yields excellent segmentation outcomes across various datasets within the SAM framework, offering valuable diagnostic data to clinicians.

Although our model has made substantial performance enhancements, there remain areas requiring further refinement. A limitation lies in the impracticality of pre-training on extensive medical datasets, attributable to time and GPU computational constraints. Moreover, the model's performance needs enhancement, particularly when segmenting image categories with ambiguous or intricate boundaries. The current model relies on manual prompts similar to SAM, which are somewhat cumbersome and inefficient. Looking forward, we intend to delve deeper into pre-training efforts and investigate automated prompt-generation strategies, such as YOLO. Such integration would facilitate automated prompt generation and enhance the model's performance.

Data availability

1.MICCAI FLARE 2022 challenge dataset: https://flare22.grand-challenge.org/Dataset/, 2.BraTs dataset: https://www.kaggle.com/datasets/dschettler8845/brats-2021-task1. The private datasets used and analysed during the current study available from the corresponding author on reasonable request.

Received: 16 June 2024; Accepted: 14 August 2024
Published online: 21 August 2024

References

1. Asgari Taghanaki, S., Abhishek, K., Cohen, J. P., Cohen-Adad, J. & Hamarneh, G. Deep semantic segmentation of natural and medical images: A review. *Artif. Intell. Rev.* **54**, 137–178. https://doi.org/10.1007/s10462-020-09854-1 (2021).
2. Smistad, E., Falch, T. L., Bozorgi, M., Elster, A. C. & Lindseth, F. Medical image segmentation on GPUs—A comprehensive review. *Med. Image Anal.* **20**, 1–18. https://doi.org/10.1016/j.media.2014.10.012 (2015).
3. Petitjean, C. & Dacher, J.-N. A review of segmentation methods in short axis cardiac MR images. *Med. Image Anal.* **15**, 169–184. https://doi.org/10.1016/j.media.2010.12.004 (2011).
4. Qureshi, I. *et al.* Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Inf. Fusion* **90**, 316–352. https://doi.org/10.1016/j.inffus.2022.09.031 (2023).
5. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/TPAMI.2016.2572683 (2015).
6. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical image Computing and Computer-assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18* 234–241. https://doi.org/10.1007/978-3-319-24574-4_28 (Springer) (2015).
7. Khadidos, A., Sanchez, V. & Li, C.-T. Weighted level set evolution based on local edge features for medical image segmentation. *IEEE Trans. Image Process.* **26**, 1979–1991. https://doi.org/10.1109/TIP.2017.2666042 (2017).
8. Kirillov, A. *et al.* Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026. https://doi.org/10.48550/arXiv.2304.02643 (2023).
9. Ma, J. *et al.* Segment anything in medical images. *Nat. Commun.* **15**, 654. https://doi.org/10.1038/s41467-024-44824-z (2024).
10. Mazurowski, M. A. *et al.* Segment anything model for medical image analysis: An experimental study. *Med. Image Anal.* **89**, 102918. https://doi.org/10.1016/j.media.2023.102918 (2023).
11. Shi, P. *et al.* Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. *Diagnostics* **13**, 1947. https://doi.org/10.3390/diagnostics13111947 (2023).
12. Roy, S. *et al.* Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model. Preprint at arXiv:2304.05396. https://doi.org/10.48550/arXiv.2304.0539 (2023).
13. Wu, J. *et al.* Medical sam adapter: Adapting segment anything model for medical image segmentation. Preprint at arXiv:2304.12620. https://doi.org/10.48550/arXiv.2304.12620 (2023).
14. Zhang, K. & Liu, D. Customized segment anything model for medical image segmentation. Preprint at arXiv:2304.13785. https://doi.org/10.48550/arXiv.2304.13785 (2023).
15. Paluru, N. *et al.* Anam-Net: Anamorphic depth embedding-based lightweight CNN for segmentation of anomalies in COVID-19 chest CT images. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 932–946. https://doi.org/10.1109/TNNLS.2021.3054746 (2021).
16. Wang, G. *et al.* DeepIGeoS: A deep interactive geodesic framework for medical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 1559–1572. https://doi.org/10.1109/TPAMI.2018.2840695 (2018).
17. Wang, S. *et al.* CT male pelvic organ segmentation using fully convolutional networks with boundary sensitive representation. *Med. Image Anal.* **54**, 168–178. https://doi.org/10.1016/j.media.2019.03.003 (2019).
18. Sun, Y. *et al.* MSCA-Net: Multi-scale contextual attention network for skin lesion segmentation. *Pattern Recogn.* **139**, 109524. https://doi.org/10.1016/j.patcog.2023.109524 (2023).
19. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211. https://doi.org/10.1038/s41592-020-01008-z (2021).
20. Liu, H., Wang, H., Wu, Y. & Xing, L. Superpixel region merging based on deep network for medical image segmentation. *ACM Trans. Intell. Syst. Technol. (TIST)* **11**, 1–22. https://doi.org/10.1145/3386090 (2020).
21. Hatamizadeh, A. *et al.* Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 574–584. https://doi.org/10.1109/WACV51458.2022.00181 (2022).
22. Hatamizadeh, A. *et al.* Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, 272–284. https://doi.org/10.1007/978-3-031-08999-2_22 (Springer, 2021).
23. Chen, J. *et al.* Transunet: Transformers make strong encoders for medical image segmentation. Preprint at arXiv:2102.04306. https://doi.org/10.48550/arXiv.2102.04306 (2021).
24. Wu, J. *et al.* Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning*, vol. 227 of *Proceedings of Machine Learning Research*, 1623–1639. https://doi.org/10.48550/arXiv.2211.00611 (PMLR, 2024).
25. Zhao, X. *et al.* Fast segment anything. Preprint at arXiv:2306.12156. https://doi.org/10.48550/arXiv.2306.12156 (2023).
26. Zhang, C. *et al.* Faster segment anything: Towards lightweight sam for mobile applications. https://doi.org/10.48550/arXiv.2306.14289 (2023).
27. Xiong, Y. *et al.* Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16111–16121. https://doi.org/10.48550/arXiv.2312.00863 (2024).

Acknowledgements
This work was supported by Major Science and Technology Innovation Project of Hangzhou under Grant (No. 2022AIZD010).

Author contributions
G.D. curated data, developed software, and drafted the manuscript. Z.W. led conceptualization, methodology, and edited the manuscript. Y.S., H.S., and H.C. conducted analysis and validation. L.L. reviewed and edited the manuscript. Y.C. managed funding and project oversight, and approved the final manuscript. All authors reviewed the manuscript.

Competing interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Additional information
Correspondence and requests for materials should be addressed to Z.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.