S-SAM: SVD-based Fine-Tuning of Segment Anything Model for Medical Image Segmentation

Jay N. Paranjape[1]✉, Shameema Sikder[2,3], S. Swaroop Vedula[3], and Vishal M. Patel[1]

[1] Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, USA
jparanj1@jhu.edu
[2] Wilmer Eye Institute, The Johns Hopkins University, Baltimore, USA
[3] Malone Center for Engineering in Healthcare, The Johns Hopkins University, Baltimore, USA

Abstract. Medical image segmentation has been traditionally approached by training or fine-tuning the entire model to cater to any new modality or dataset. However, this approach often requires tuning a large number of parameters during training. With the introduction of the Segment Anything Model (SAM) for prompted segmentation of natural images, many efforts have been made towards adapting it efficiently for medical imaging, thus reducing the training time and resources. However, these methods still require expert annotations for every image in the form of point prompts or bounding box prompts during training and inference, making it tedious to employ them in practice. In this paper, we propose an adaptation technique, called S-SAM, that only trains parameters equal to 0.4% of SAM's parameters and at the same time uses simply the label names as prompts for producing precise masks. This not only makes tuning SAM more efficient than the existing adaptation methods but also removes the burden of providing expert prompts. We call this modified version S-SAM and evaluate it on five different modalities including endoscopic images, x-ray, ultrasound, CT, and histology images. Our experiments show that S-SAM outperforms state-of-the-art methods as well as existing SAM adaptation methods while tuning a significantly less number of parameters. We release the code for S-SAM at https://github.com/JayParanjape/SVDSAM.

Keywords: Blackbox Adaptation · Prompted Segmentation.

1 Introduction

Image segmentation is a fundamental task in medical image analysis. Various deep learning (DL) based algorithms have been proposed in the literature that are able to perform remarkably well on a wide variety of modalities including CT, MRI, X-ray, ultrasound and endoscopy for segmenting organs, tumors, and

tissues [2,32]. However, most of the well-performing models inherently have a large number of parameters, thus increasing the resources needed to train them every time for a new dataset or a new modality. This problem has been tackled for natural images to some extent by foundational models which are trained on billions of data points, exhibit an innate understanding of a given task, and are able to generalize well, given suitable user prompts. Notable examples include CLIP [22] and ALIGN [11] for image classification and open set image-text understanding, and the recently proposed Segment Anything Model (SAM) [13] for segmentation. As such, it was hoped that medical image segmentation could also be addressed using SAM. However, SAM is not trained on medical images and various studies show its inability to generalize well on medical data [9,21]. In order to utilize the power of SAM on medical images, there have been various efforts to adapt it efficiently [19,33,24,18]. However, it is not feasible to provide expert-level prompts for every image during training or inference, as is required by these adaptation methods since this process is time-consuming for experts. In our work, we propose S-SAM - an adaptation of SAM that only expects the name of the class of interest as a prompt, thus eliminating the need for expert-level prompts. This property makes it more suitable for usage in medical systems while also retaining the promptable nature of SAM. Furthermore, we show that our approach is significantly more efficient and requires training far fewer parameters than existing methods. In the proposed approach, we tune the singular values of the weight matrices in SAM's image encoder. Singular values define how important each visual feature in the image is to the activation of a given layer of the model [29] and hence, modifying these during training allows S-SAM to correctly adapt SAM for a given task.

In summary, the main contributions of our paper are as follows. (1) We propose a novel adaptation of SAM, called S-SAM, that can perform text-prompted segmentation of medical images. S-SAM expects prompts as simple as the name of the class of interest to produce precise masks, thus removing the requirement of expert-level prompts. (2) We develop a technique for training S-SAM that makes it far more efficient than existing adaptation methods. In comparison with the original SAM, the number of trainable parameters for S-SAM are only 0.4%. (3) Extensive experiments are conducted on five publicly available datasets of different modalities where we obtain the SOTA performance.

2 Related Work

While SAM is a powerful tool for natural image segmentation, it needs to be adapted to perform well in the medical domain. Hence, much research has been conducted to harness the power of SAM's encoders and decoders for medical images. One of the first efforts in this direction was MedSAM [19], where the authors finetune SAM for a huge corpus of medical images and provide support for point and bounding box-based prompts. While this approach finetunes all the parameters of SAM, the Medical SAM Adapter [33] introduces learnable adapter layers for all the encoder and decoder blocks. These are low-rank approximations

that allow learning the domain shift while freezing all other parameters of SAM. While this method significantly brings down the GPU requirement from 1024 (for SAM) to 4, it can still be considered compute-intensive. SonoSAM [24] tunes the mask decoder and prompt encoder of SAM for training on sonography images. SAMUS [18], on the other hand, adapts SAM for ultrasound images by training an additional CNN-based image encoder and fusing it with SAM's encoder. We refer readers to a comprehensive survey of SAM-adaptation methods [15] for more details. However, these approaches still require experts to manually provide precise point prompts or bounding box prompts for every image during training and testing, which is tedious. Hence, an adaptation of SAM that minimizes expert involvement would be highly beneficial for medical applications. In light of this, SAMed [35] was introduced, which added trainable low-rank adaptation layers to SAM's image encoder while keeping the rest of the encoder weights frozen. In addition, the user-defined prompts were replaced with a default prompt to eliminate expert involvement. Similarly, AutoSAM [26] replaces the user-defined prompt with the image itself, and trains the prompt encoder to produce good prompt embeddings with the image input. However, these techniques remove the promptable nature of SAM. To tackle this, AdaptiveSAM [21] introduces text-prompted segmentation which allows the use of label names as the prompt. To perform this, they tune the biases of the encoder network and keep the decoder fully trainable while also adding a text affine layer, producing good results for surgical scene segmentation. Using text prompts can be considered as requiring no expert involvement since simply the label names can serve as a valid text prompt. Hence, in S-SAM, we also perform text-prompted segmentation. However, our method does not require the mask decoder of SAM to be tuned. Hence, it is more efficient than AdaptiveSAM. Furthermore, by tuning the singular values instead of biases, our model does a better job of learning the domain shift. In summary, an appropriate adaptation method for SAM should minimize expert involvement. It should allow for some form of prompting to facilitate interactivity as needed. Finally, it should be as efficient as possible. Our method satisfies all of these requirements, as shown in Table 1.

Table 1: Comparison of different adaptation methods of SAM present in the literature. Here, expert intervention means providing expert prompts for every image, including points, bounding boxes or masks. We exclude text prompts with label names from this category since experts are not required to provide the label name prompts for every image, but only once for the entire dataset.

Method	Expert Intervention Not Required	Promptable	Training Decoder Not Required
MedSAM [19]	✗	✓	✗
Medical SAM Adapter [33]	✗	✓	✗
SonoSAM [24]	✗	✓	✗
SAMUS [18]	✗	✓	✓
SAMed [35]	✓	✗	✗
AutoSAM [26]	✓	✗	✓
AdaptiveSAM [21]	✓	✓	✗
S-SAM (Ours)	✓	✓	✓

3 Proposed Method

SVD-based Tuning: The image encoder in SAM is a chain of N blocks, each of which has a multi-head attention module followed by a Multi Layer Perceptron (MLP), which perform the following computation

$$qkv = (W^n_{qkv})x + b^n_{qkv}, \quad o_n = (W^n_M)x + b^n_M, \tag{1}$$

where q, k and v denote the query, key and value associated with the multi-head attention and M denotes the MLP layer. o denotes the output of the block and x denotes the input to a module. Here, n is used to index the number of blocks (N). Note that $W \in \mathbb{R}^{D\times K}$ and $b \in \mathbb{R}^{D\times 1}$ denote the weights and biases of the respective modules. Here, D represents the dimension of the input to the module (multi-head attention or MLP) and K represents the output dimension of the respective module. These weights are primarily responsible for learning how to encode natural images after being extensively trained on them. To adapt them to the medical domain, we propose tuning W as follows:

$$W \leftarrow UReLU(A \odot \Sigma + B)V^T, \quad \text{where } W = U\Sigma V^T, \tag{2}$$

where $W = U\Sigma V^T$ denotes the Singular Value Decomposition (SVD) of the weight matrix W and $\odot$ represents the element-wise multiplication. A and B are matrices with the same shape as Σ with non-diagonal entries as 0. In other words, we compute the singular values of any weight matrix in the image encoder and learn a non-linear transformation over them. Since A is multiplied elementwise, it represents scaling of the singular values while B is added and thus represents shifting of the singular values. Finally, to maintain the positive semidefinite nature of the Σ matrix, we apply a ReLU operation. This operation is illustrated in Figure 2 (c).

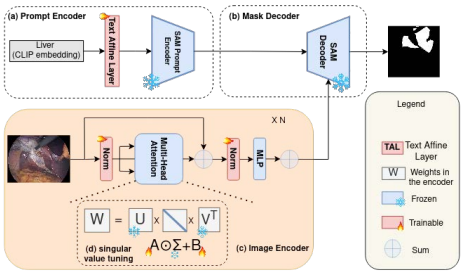

Fig. 1: S-SAM Architecture. The image encoder weights are modified by performing a transformation over their singular values. Other trainable parameters include the layernorms and positional embeddings in the encoder and the Text Affine Layer (TAL). Everything else is frozen and initialized with SAM's pretrained checkpoint.

S-SAM - Architecture: An overview of S-SAM's architecture can be found in Figure 1. S-SAM consists of an image encoder, prompt encoder and the mask decoder. An additional module called the Text Affine Layer (TAL) is also added before the prompt encoder. The inputs to S-SAM include an image and a text prompt (which can be the label name itself) and the output is a mask over the region described by the text. The image encoder of S-SAM is SAM's original image encoder, albeit with all the weights modified as described in the previous subsection. The image is fed into this modified encoder which

outputs the image embeddings. Similarly, for the text prompt, we first obtain the CLIP embedding. However, both CLIP and SAM were trained on natural images and hence might not capture the medical labels correctly. Therefore, we pass the CLIP embedding through TAL (a learnable one-layer MLP), before passing it to the prompt encoder, which outputs the prompt embeddings. Finally, we use the mask decoder to fuse the image and prompt embeddings to generate the mask of interest.

In this setup, all the weights of the image encoder, prompt encoder and mask decoder are initialized with SAM's pre-trained checkpoint. The non-zero elements of A corresponding to every weight matrix are initialized with one and B is initialized with zeros. During training, CLIP, the mask decoder and the prompt encoder are completely frozen. All the weights of the image encoder are frozen and only the transform parameters A and B are learnable. In addition, positional embeddings are trainable to allow training with smaller resolutions and layernorm layers are trainable to better adapt to the new domain. The positional embeddings in SAM's checkpoint expect the input image size to be 1024×1024. Hence, many adaptation techniques upsample the images to this scale. However, this significantly adds on to the memory requirements of these approaches. We replace SAM's positional embeddings with learnable embeddings that can facilitate training with lower image sizes. To retain the information contained in SAM, we initialize these by performing an AveragePool operation over the embedding weights present in SAM's checkpoint to bring them to the required size. Finally, the TAL is also trainable.

Comparison with LoRA: Low Rank Adaptation (LoRA) is a highly effective technique for fine-tuning large language models for various applications [10,5,34]. This concept involves adding a product of two low-rank matrices $X \in \mathbb{R}^{D\times r}$ and $Y \in \mathbb{R}^{r\times K}$ to W as follows: $W \leftarrow W + XY$. However, this may lead to underfitting if the learnt subspace is smaller than required [35]. Thus, the effectiveness of this method largely depends on the rank of the approximated matrices. In S-SAM, all the singular values are tuned resulting in a full-rank computation, thus avoiding this problem. Furthermore, LoRA involves fine-tuning all the parameters of the low-rank matrices while S-SAM only tunes the singular values. This makes our approach more effi-

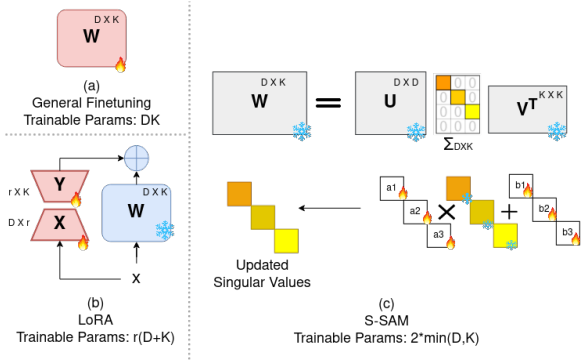

Fig. 2: Comparison of different fine-tuning methods. (a) Naive fine-tuning (b) LoRA (c) Our approach only tunes the singular values and is even more efficient than LoRA.

cient than LoRA. Assuming W is $D \times K$, the rank of W is given by $\min(D, K)$. In LoRA, the learnt matrices $X \in \mathbb{R}^{DXr}$ and $Y \in \mathbb{R}^{r \times K}$ have rank $r \ll \min(D, K)$. This makes the number of trainable parameters equal to $r(D + K)$ as shown in Figure 2. On the other hand, S-SAM only tunes the singular values of W by learning the scale A and shift B parameters. Hence, the number of trainable parameters amounts to $2 \times \min(D, K)$, which is lesser than LoRA.

4 Experimental Results

Datasets: We evaluate S-SAM on the following five medical imaging datasets corresponding to different modalities. (i) CholecSeg8k [8] consists of endoscopic surgery images with 12 classes of interest including surgical instruments, organs and tissues. (ii) The abdominal ultrasound dataset [31] consists of simulated ultrasound images and has 8 classes of interest denoting different organs and bones. The testing data consists of a mix of real and synthetic images. (iii) ChestXDet [17] has x-ray images with 13 classes of interest representing malignancies in the chest region. (iv) The LiTS Dataset [20] consists of Computed Tomography (CT) images with the classes of interest being the liver and the tumor region. S-SAM takes in a 2D image as an input. Hence, we use its derived dataset having 2D slices found at [20]. (v) The GLAS challenge dataset [28] comprises of histology images. There is only one class of interest, namely the glands, which needs to be segmented. The rest of the experimental setup is outlined in the supplementary.

Results: In clinical practice, a blank mask, corresponds to the label of interest that is not present in the image. Note that a blank mask is a valid prediction. For example, if a CT scan of a normal liver is queried with the text prompt "Tumor", it should output an empty mask. In the classical definition of DICE Score (DSC), such cases have undefined score and hence, ignored. Hence, for each of the datasets, we evaluate S-SAM using the Dice Score as defined by Rahman et al. [23] who give a DSC of 1 to cases where the predicted mask and the label are both completely blank.

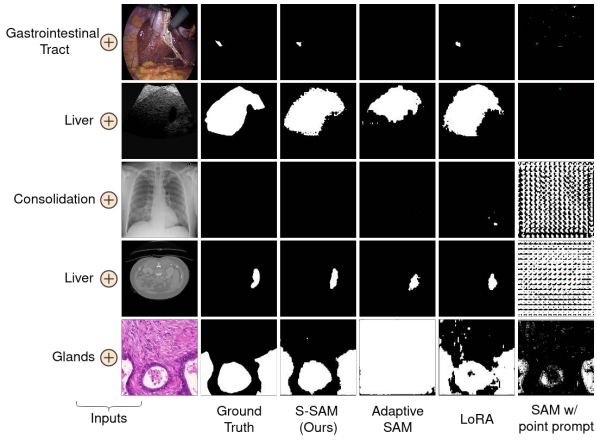

Fig. 3: A qualitative comparison among different methods. From the top, the rows represent CholecSeg8k, Ultrasound, ChestXDet, LiTS, and GLAS, respectively. The green dot in the last column denotes the point prompt used to query SAM.

Table 2: Results on ChoecSeg8k. Some methods group certain labels into one category and report the groupwise results, denoted by multi-columns numbers.

GB-Gall Bladder, AW-Abdominal Wall, GT-GI Tract, CD-Cystic Duct, LHEC-L Hook Electrocautery, HV-Hepatic Vein, CT-Connective Tissue, LL-Liver Ligament.													
Method	Object wise DSC												
	Fat	Liver	GB	AW	GT	Grasper	LHEC	Blood	HV	CT	LL	CD	Avg.
Traditional DL Methods													
U-Net[25]	0.87	0.52	0.40	0.73	0.26	0.52		0.08					0.48
UNetR [7]	0.88	0.74	0.42	0.76	0.35	0.71		0					0.55
TransUNet [3]	0.83	0.43	0.77	0.35	0.43	0.70	0.55	0.61	0.82	0.57	0.72	0.64	0.62
MedT [30]	0.81	0.39	0.56	0.34	0.25	0.48	0.71	1	0.70	0.69	0	0.89	0.57
SAM based methods													
SAM w/ text prompt [13]	0.05	0	0.02	0	0	0.01	0.04	0.01	0.14	0.01	0.14	0.01	0.04
SAM w/ point prompt [13]	0.17	0.23	0.07	0.30	0.10	0.22	0.63	1	0.70	0.43	1	1	0.49
SAM with full finetuning [13]	0	0	0.02	0	0.09	0.13	0.38	0.91	0.7	0.38	1	0.91	0.38
MedSAM [19]	0	0	0.02	0	0.08	0.15	0.46	1	0.69	0.38	1	1	0.40
SAMed [35]	0	0	0.03	0	0.13	0.19	0.46	1	0.69	0.38	1	1	0.41
AdaptiveSAM [21]	0.85	0.71	0.37	0.80	0.10	0.20	0.70	1	0.70	0.38	1	1	0.64
Low Rank Adaptation of SAM [10]	0.87	0.72	0.45	0.76	0.42	0.20	0.48	0.97	0.70	0.70	0.6	0.97	0.65
S-SAM (Ours)	0.89	0.71	0.51	0.73	0.43	0.30	0.63	1	0.71	0.56	1	1	**0.71**

We present quantitative results for CholecSeg8k, LiTS, and GLAS in Tables 2, 4, and 5 respectively. Due to space constraints, results on Abdominal Ultrasound and ChestXDet are in supplementary Tables 2 and 3 respectively. On average, we achieve a significant improvement of 6-7% on CholecSeg8k, Ultrasound, and GLAS over the existing state-of-the-art (SOTA), and 1% improvement on the other two datasets. However, note that our method achieves on par performance with a significantly lesser number of parameters. From the tables, we see that adapting foundation models like SAM improves performance over existing DL-based methods. Furthermore, we compare S-SAM with zero-shot SAM in the first two rows in SAM-based methods in the tables. SAM is not trained on medical images and hence performs poorly when prompted with the label name as text prompt without any finetuning. However, with a point prompt, it can still segment to some extent. Similarly, MedSAM also performs well on certain objects, but not as well as adaptation methods, showing the requirement of tuning on new datasets. We also observe that full finetuning of SAM overfits to the data, resulting in a lower test performance. Finally, we show improvements over various SAM-based adaptation methods. A comparison based on the number of parameters is provided in Figure 4, which shows that S-SAM requires a significantly lower number of parameters than these adaptation methods, while also outperforming them on all five datasets. In comparison to SAM, there is a 99.6% reduction in the number of trainable parameters while in comparison to AdaptiveSAM, there is 90% reduction. Similarly, S-SAM trains 50% lesser parameters than LoRA. All results with S-SAM have a p-value of at most 10^{-8} wrt other methods, which shows statistical significance.

Qualitative Results: In Figure 3, we present sample results of S-SAM and other methods. S-SAM is able to segment smaller objects like the GI Tract that is missed by Adaptive SAM or SAM as seen in row 1 of the Figure. In addition, S-SAM is also able to produce blank masks when a queried object of interest is not present, as seen

Table 3: Ablation analysis on the components of S-SAM.

Tuning Pos Embeds	Tuning Layernorm	TAL	Scaling	Shifting	Avg DSC
					0.04
✓					0.04
✓	✓				0.50
✓	✓	✓			0.52
✓	✓	✓	✓		0.54
✓	✓	✓		✓	0.64
✓	✓	✓	✓	✓	**0.71**

in row 3. The effectiveness of adaptation methods can be visually represented through all the cases, where zero-shot SAM produces gibberish results, unlike the adaptation methods.

Table 4: Results on LiTS

Method	Objectwise DSC		
	Liver	Tumor	Avg.
Traditional DL Methods			
UNet [25]	0.77	0.65	0.71
SegNet [1]	0.76	0.64	0.70
KiuNet [12]	0.80	0.71	0.76
DeepLab v3+ [4]	0.85	0.68	0.77
SAM based Methods			
SAM w/ text prompt [13]	0.04	0	0.02
SAM w/ point prompt [13]	0.05	0	0.03
SAM with full finetuning [13]	0.05	0.86	0.5
MedSAM [19]	0.06	0.01	0.04
SAMed [35]	0.61	0.91	0.76
AdaptiveSAM [21]	0.80	0.86	0.83
Low Rank Adaptation of SAM [10]	0.82	0.84	0.83
S-SAM (Ours)	0.85	0.83	**0.84**

Table 5: Results on GLAS.

Method	Objectwise DSC
	Glands
Traditional DL Methods	
UNet [25]	0.52
SegNet [1]	0.84
clDice [27]	0.85
PointRend [14]	0.88
TI-Loss [6]	0.88
SAM Based Methods	
SAM w/ text prompt [13]	0.01
SAM w/ point prompt [13]	0.19
SAM with full finetuning [13]	0.85
MedSAM [19]	0.21
SAMed [35]	0.65
AdaptiveSAM [21]	0.66
Low Rank Adaptation of SAM [10]	0.83
S-SAM (Ours)	**0.90**

Component-wise Ablation: S-SAM has three major modified components over SAM, namely the Text Affine Layer (TAL), scaling matrix A, and the shifting matrix B, as well as other modifications like training the positional embeddings and layernorms of the model. To assess the importance of each of these, we start with SAM and note the rise in performance on CholecSeg8k when each component is added. Results for this study are tabulated in Table 3.

The first row in the table shows SAM's zero-shot performance without modifications. Tuning only the positional embeddings of the encoder doesn't improve performance. However, tuning layernorm layers significantly boosts DSC by around 46%, consistent with domain adaptation research where norm layers contribute to domain bias [16]. Adding TAL further improves performance. Shifting and scaling, added one by one, both enhance performance, with shifting appearing more important for modeling domain shift, resulting in a 12% increase. Finally, with all components, we achieve S-SAM with the best performance.

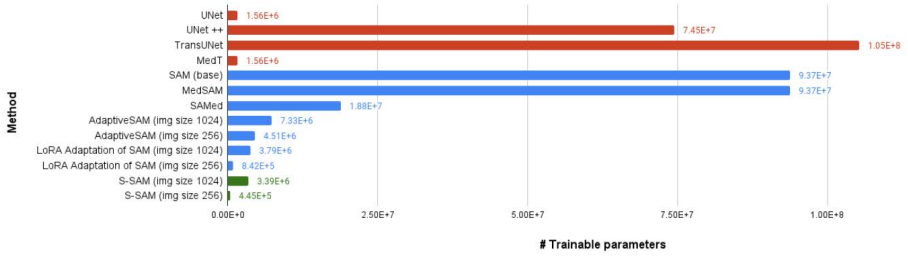

Fig. 4: A comparison among different methods based on the number of parameters trained. The red bars indicate traditional DL-based segmentation methods. Blue bars indicate SAM-based methods and green bars indicate our method. The numbers to the right of each bar denote the number of trainable parameters.

5 Conclusion

In this paper, we presented S-SAM - an efficient adaptation of SAM for medical images, which is realized by tuning the singular values of the weight matrix. We show that S-SAM performs on par or outperforms existing methods on five publicly available medical image segmentation datasets with significantly lower amount of trainable parameters and allows the use of label names as prompts. We find that S-SAM could be affected by class-size disparities, as seen from its performance on specific classes. A potential future direction towards improving S-SAM could be using additional loss functions or weighted loss functions to reduce the effect of the disparities.

6 Disclosure of Interests

This research was supported by a grant from the National Institutes of Health, USA; R01EY033065. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors have no competing interests in the paper

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation (2016)
2. Basak, H., Yin, Z.: Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19786–19797 (June 2023)
3. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation (2018)
5. Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.M., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H.T., Chen, J., Liu, Y., Tang, J., Li, J., Sun, M.: Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence **5** (2023)
6. Gupta, S., Hu, X., Kaan, J., Jin, M., Mpoy, M., Chung, K., Singh, G., Saltz, M., Kurc, T., Saltz, J., Tassiopoulos, A., Prasanna, P., Chen, C.: Learning topological interactions for multi-class medical image segmentation (2022)
7. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1748–1758 (2022)
8. Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S.: Cholecseg8k: A semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80 (2020)
9. Hu, C., Xia, T., Ju, S., Li, X.: When sam meets medical images: An investigation of segment anything model (sam) on multi-phase liver tumor segmentation (2023)

10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021)
11. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision (2021)
12. Jose, J.M., Sindagi, V., Hacihaliloglu, I., Patel, V.M.: Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations (2020)
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
14. Kirillov, A., Wu, Y., He, K., Girshick, R.: Pointrend: Image segmentation as rendering (2020)
15. Lee, H.H., Gu, Y., Zhao, T., Xu, Y., Yang, J., Usuyama, N., Wong, C., Wei, M., Landman, B.A., Huo, Y., Santamaria-Pang, A., Poon, H.: Foundation models for biomedical image segmentation: A survey (2024)
16. Li, Y., Wang, N., Shi, J., Liu, J., Hou, X.: Revisiting batch normalization for practical domain adaptation (2016)
17. Lian, J., Liu, J., Zhang, S., Gao, K., Liu, X., Zhang, D., Yu, Y.: A structure-aware relation network for thoracic diseases detection and segmentation (2021)
18. Lin, X., Xiang, Y., Zhang, L., Yang, X., Yan, Z., Yu, L.: Samus: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation (2023)
19. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images (2023)
20. Maranhão, A.: https://www.kaggle.com/datasets/andrewmvd/lits-png, https://www.kaggle.com/datasets/andrewmvd/lits-png
21. Paranjape, J.N., Nair, N.G., Sikder, S., Vedula, S.S., Patel, V.M.: Adaptivesam: Towards efficient tuning of sam for surgical scene segmentation (2023)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
23. Rahman, A., Valanarasu, J., Hacihaliloglu, I., Patel, V.M.: Ambiguous medical image segmentation using diffusion models. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11536–11546. IEEE Computer Society, Los Alamitos, CA, USA (jun 2023), https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01110
24. Ravishankar, H., Patil, R., Melapudi, V., Bhatia, P., Taha, K.H., Annangi, P.: Sonosam – segment anything on ultrasound images (2023)
25. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. vol. 9351 (2015)
26. Shaharabany, T., Dahan, A., Giryes, R., Wolf, L.: Autosam: Adapting sam to medical images by overloading the prompt encoder (2023)
27. Shit, S., Paetzold, J.C., Sekuboyina, A., Ezhov, I., Unger, A., Zhylka, A., Pluim, J.P.W., Bauer, U., Menze, B.H.: clDice - a novel topology-preserving loss function for tubular structure segmentation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2021), https://doi.org/10.1109%2Fcvpr46437.2021.0162
28. Sirinukunwattana, K., Pluim, J.P.W., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., Böhm, A., Ronneberger, O., Cheikh, B.B., Racoceanu, D., Kainz, P., Pfeiffer, M., Urschler, M., Snead,

D.R.J., Rajpoot, N.M.: Gland segmentation in colon histology images: The glas challenge contest (2016)
29. Turk, M., Pentland, A.: Face recognition using eigenfaces. In: Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 586–591 (1991)
30. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. pp. 36–46. Springer International Publishing, Cham (2021)
31. Vitale, S., Orlando, J., Iarussi, E., Larrabide, I.: Improving realism in patient-specific abdominal ultrasound simulation using cyclegans. International Journal of Computer Assisted Radiology and Surgery (07 2019)
32. Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., Nandi, A.K.: Medical image segmentation using deep learning: A survey. IET Image Processing **16**(5), 1243–1267 (2022), https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/ipr2.12419
33. Wu, J., Zhang, Y., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation (2023)
34. Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., Hu, X.: Harnessing the power of llms in practice: A survey on chatgpt and beyond (2023)
35. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation (2023)