Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation

Junde Wu[1,2,4], Rao Fu[1], Yu Zhang[1], Huihui Fang[2], Yuanpei Liu[3], Zhaowei Wang[4], Yanwu Xu[2] (✉), and Yueming Jin[5] (✉)

[1] BoomSlang Med. Tech.
[2] Pazhou Lab
[3] The University of Hong Kong
[4] Baidu Research
[5] National University of Singapore
work in progress

Abstract. The Segment Anything Model (SAM) has recently gained popularity in the field of image segmentation. Thanks to its impressive capabilities in all-round segmentation tasks and its prompt-based interface, SAM has sparked intensive discussion within the community. It is even said by many prestigious experts that image segmentation task has been "finished" by SAM. However, medical image segmentation, although an important branch of the image segmentation family, seems not to be included in the scope of Segmenting "Anything". Many individual experiments and recent studies have shown that SAM performs subpar in medical image segmentation. A natural question is how to find the missing piece of the puzzle to extend the strong segmentation capability of SAM to medical image segmentation. In this paper, instead of fine-tuning the SAM model, we propose Med SAM Adapter, which integrates the medical specific domain knowledge to the segmentation model, by a simple yet effective adaptation technique. Although this work is still one of a few to transfer the popular NLP technique Adapter to computer vision cases, this simple implementation shows surprisingly good performance on medical image segmentation. A medical image adapted SAM, which we have dubbed Medical SAM Adapter (MSA), shows superior performance on 19 medical image segmentation tasks with various image modalities including CT, MRI, ultrasound image, fundus image, and dermoscopic images. MSA outperforms a wide range of state-of-the-art (SOTA) medical image segmentation methods, such as nnUNet, TransUNet, UNetr, MedSegDiff, and also outperforms the fully fine-turned MedSAM with a considerable performance gap. Code will be released at: https://github.com/WuJunde/Medical-SAM-Adapter.

1 Introduction

Very recently, the Segmentation Anything Model (SAM) [21] has gained massive attention as a powerful and general vision segmentation model capable of generating various and fine-grained segmentation masks conditioned by the user prompt. Despite its strong performance over natural images, many recent studies

show [9, 16, 31] that it exhibits subpar performance on medical image segmentation. Medical image segmentation, however, is a very important branch of image segmentation that is a fundamental step in many clinical applications such as diagnosis, surgical planning, and image-guided surgery. This is definitely a significant part of the image segmentation family that an expected segmentation "anything" model should not ignore.

The main reason for SAM's failure over medical images is due to the lack of training data. Although SAM established a sophisticated and efficient data engine in the training, they collected few cases for medical usage. In this paper, we attempt to expand SAM towards prompt-based medical image segmentation with minimum effort. Technically, we choose to fine-tune the pre-trained SAM using a parameter-efficient fine-tuning (PEFT) technique called Adaption [18]. Adaption is a popular and widely-used technology in natural image processing (NLP) to fine-tune the fundamental pre-trained model for specific usage. The main idea is to insert several parameter-efficient Adapter modules into the original fundamental model, and then only adjust the Adapter parameter while leaving all pre-trained parameters frozen. In this work, we make several modifications for our specific medical usage. Surprisingly, this simple strategy yields excellent performance in our validation experiments. We will further explain why we chose this specific technique route using the following questions and answers:

1. Why we need SAM for medical image segmentation?

We believe that interactive (prompt-based) segmentation is the paradigm of all segmentation tasks. The prompt, which determines the granularity of the expected results, is necessary for zero-shot segmentation and should be provided by the users. For instance, in the case of medical images, different targets on a fundus image, such as vessels, optic disc, optic cup, and macula, may need to be segmented from a single fundus image, depending on different requirements and usage. SAM provides an excellent framework for interacive segmentation, making it a perfect starting point towards prompt-based medical image segmentation.

2. Why fine-tuning?

SAM's pre-trained model has been trained over the largest segmentation dataset worldwide through a well-designed data engine. Such a model is valuable for medical image segmentation, as many studies have shown that pretraining on natural images is also beneficial for medical image segmentation [30, 38], at least in terms of convergence speed.

3. Why PEFT and Adaption?

PEFT has proven to be an efficient strategy for fine-tuning a large, fundamental model for a specific usage [41]. Compared to full fine-tuning, it keeps most of the parameters frozen and learns significantly fewer parameters, often less than 5% of the total. This enables efficient learning with faster updates. Studies have also shown that PEFT approaches work better than full fine-tuning as they avoid catastrophic forgetting and generalize better to out-of-domain scenarios, especially in low-data regimes [41]. Among all PEFT strategies, Adaption [18] stands out as an effective tool for fine-tuning large fundamental vision models for downstream tasks, not only in NLP but also in computer vision. Recent

studies have shown that Adaption can be easily adopted in various downstream computer vision tasks [4, 17]. Therefore, we believe Adaption is the most fitting technique for carrying SAM to the medical domain.

Our contributions can be summarized as follows:

- We have extended the capabilities of the popular and powerful SAM model to medical domain, which is a significant step towards the ultimate goal of "Segment Anything".
- To the best of our knowledge, we are the first to propose the adaptation approach for the general medical image segmentation. We consider the domain-specific knowledge when designing the adapter, such as high dimensionality (3D) of medical data, and unique setting of visual prompts, such as click and bbox prompts for the decoder.
- We have evaluated our proposed MSA model on 19 medical image segmentation tasks with different image modalities including MRI, CT, fundus image, ultrasound image, and dermoscopic images. Our results demonstrate that MSA outperforms the previous state-of-the-art methods by a considerable margin.

2 Method

2.1 Preliminary: SAM architecture

To begin with, we provide an overview of the SAM architecture before discussing our modifications. SAM comprises three main components: an image encoder, a prompt encoder, and a mask decoder. The image encoder is based on a standard Vision Transformer (ViT) pre-trained by MAE. Specifically, we use the ViT-H/16 variant, which employs 14×14 windowed attention and four equally-spaced global attention blocks. The output of the image encoder is a 16× downsampled embedding of the input image. The prompt encoder can be either sparse (points, boxes, text) or dense (masks). In this paper, we focus only on the sparse encoder, which represents points and boxes as positional encodings summed with learned embeddings for each prompt type. The mask decoder is a Transformer decoder block modified to include a dynamic mask prediction head. SAM uses two-way cross-attention, one for prompt-to-image embedding and another for image-to-prompt embedding in each block to learn the interaction between the prompt and image embeddings. After running two blocks, SAM upsamples the image embedding, and an MLP maps the output token to a dynamic linear classifier, which predicts the target mask of the given image.

2.2 MSA architecture

Our goal is to fine-tune the SAM architecture for medical image segmentation tasks to enhance its medical ability. Instead of fully turning all parameters, we kept the pre-trained SAM parameters frozen and inserted an Adapter module at specific positions in the architecture. The Adapter is a bottleneck model that
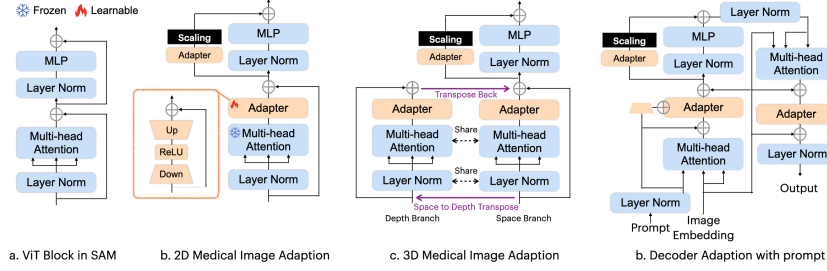
Fig. 1: MSA architecture. Note that we did not show the prompt self-attention module and mask prediction head in the decoder.

sequentially uses a down-projection, ReLU activation, and up-projection. An illustration is shown in Fig. 2 (b). The down-projection compresses the given embedding to a smaller dimension using a simple MLP layer. The up-projection expands the compressed embedding back to its original dimension using another MLP layer.

In the SAM encoder, we have deployed two adapters for each ViT block. For a standard ViT block (as shown in Fig. 2 (a)), we have placed the first Adapter after the multi-head attention and before the residue connection (as shown in Fig. 2 (b)), and the second Adapter in the residue path of the MLP layer after the multi-head attention. Immediately after the second Adapter, we have scaled the embedding with a scale factor s following [4].

In the SAM decoder, we have deployed three adapters for each ViT block. The first Adapter is deployed after the prompt-to-image embedding multi-head cross attention with a residue addition of the prompt embedding. We have made a small modification on the Adapter to integrate the prompt information into the module. Specifically, we have used another down projection to compress the prompt embedding, and added it into the Adapter over the embedding before ReLU activation. This modification helps the Adapter to turn the parameter conditioned on the prompt information and be more flexible and general to different modalities and downstream tasks. The second Adapter in the decoder is deployed in exactly the same way as in the encoder, to adapt the MLP-enhanced embedding. The third Adapter is deployed after the residue connection of the image embedding-to-prompt cross-attention. Another residue connection and layer normalization are connected after the adaption to output the final results. Note that we have only deployed the Adapter in the first block of the two blocks of the Decoder. The second block and the mask prediction head are fully turned over the given data.

Another challenge in adapting SAM to medical image segmentation is the difference in image dimension. Unlike 2D images, many medical scans are 3D,

such as MRI and CT. These 3D modalities are crucial in clinical usage, as physicians often need to recognize the correlation between slices to make decisions. Although SAM can be applied to each slice of a volume to get the final segmentation, it does not consider the correlation in the depth dimension. Many previous studies have shown that depth correlation is essential for 3D medical image segmentation [14,15,39]. To address this issue, we propose a novel adaptation method inspired by the image-to-video adaptation, with some modifications. The specific architecture is shown in Fig. 2 (c). In each block, we split the attention operation into two branches: the space branch and the depth branch. For a given 3D sample with depth D, we send D x N x L to the multi-head attention in the space branch, where N is the number of embeddings, and L is the length of the embedding. Here, D is the number of operations, and the interaction is applied over N x L to learn and abstract space correlations as embeddings. In the depth branch, we first transpose the input matrix to obtain N x D x L and then send it to the same multi-head attention. Although we use the same attention mechanism, the interaction is applied over D x L. In this way, depth correlations are learned and abstracted. Finally, we transpose the results from the depth branch back to their original shape and add them to the results of the space branch.

2.3   Training Strategy

Encoder Pretrainig Following the idea of SAM, we conduct pre-training on the model encoder using medical images. We use a mixture of four medical image datasets for this pretraining, including the RadImageNet [25] dataset, a large-scale collection containing 1.35 million radiology images (CT, MRI, US) covering a wide range of organs such as the ankle/foot, brain, hip, knee, shoulder, spine, abdomen, pelvis, chest, pelvis, and thyroid. The EyePACSp dataset [5] contains 88,702 color fundus images captured under various conditions by various devices at multiple primary care sites. The BCN-20000 [8] and HAM-10000 [33] datasets, containing approximately 30,000 dermoscopic images with melanoma or nevus on the images, are also used. All of these datasets are publicly available.

Instead of the MAE pretraining used in SAM, we use a combination of several self-supervised learning methods for pre-training. The first two are Contrastive Embedding-Mixup (e-Mix) and Shuffled Embedding Prediction (ShED), following [32]. e-Mix is a contrastive objective that additively mixes a batch of original input embeddings, weighting them with different coefficients. It then trains an encoder to produce a vector for a mixed embedding that is close to the original inputs' embeddings in proportion to their mixing coefficients. ShED shuffles a fraction of embeddings and trains the encoder with a classifier to predict which embeddings were perturbed. We also use a Masked Autoencoder (MAE) following the original implementation of SAM, which masks a given fraction of input embeddings and trains models to reconstruct them.

Training with Prompt We adapt the SAM with prompt over our new datasets. This process is basically the same as that in SAM but with modifications as follows:

For the click prompt, positive clicks indicate foreground regions, and negative clicks indicate background regions. We use a combination of random and iterative click sampling strategies for training with this prompt. Specifically, we first use random sampling for initialization, and then a few clicks are added using the iterative sampling procedure. The iterative sampling strategy resembles the interaction with a real user because, in practice, every new click is placed in the erroneous region of a prediction produced by a network using the set of previous clicks. We follow [22] to generate the random sampling and [24] to simulate the iterative sampling.

We use a different text prompt training strategy with SAM. In SAM, the authors used the image embedding of the target object crop produced by CLIP as the image embedding close to its corresponding text description or definition in CLIP. However, since CLIP is barely trained on medical image datasets, it can hardly relate the organs/lesions on the image with the corresponding text definition. Instead, we first randomly generate several free texts containing the definition of the target (i.e., optic disc, brain tumor) as the keyword from Chat-GPT, and then extract the embedding of the text using CLIP as the prompt for training. One free text could contain multiple targets, in which case we supervise the model with all their corresponding masks.

3   Experiments

3.1   Dataset

We conducted experiments on five different medical image segmentation datasets, categorized into two types. The first type tested the overall segmentation performance of the model, comparing it to commonly used medical segmentation baselines and state-of-the-art (SOTA) methods. To verify the general segmentation performance, we chose abdominal multi-organ segmentation, one of the most representative cases in medical image segmentation and a widely-used benchmark for related studies. We used two datasets, AMOS2022 [20] and BTCV [11], both publicly available and annotated with sixteen and twelve anatomies, respectively.

The other four tasks were used to verify the model's generalization to different modalities and kinds of tasks, including optic disc and optic cup segmentation over fundus images, brain tumor segmentation over brain MRI images, thyroid nodule segmentation over ultrasound images, and melanoma or nevus segmentation from dermoscopic images. For the fundus image segmentation, we conducted experiments on two publicly released datasets: REFUGE [10, 28] and RIGA [29]. For brain tumor segmentation, we conducted experiments on the publicly released BraTs dataset [2, 12, 26]. For thyroid nodule segmentation, we conducted experiments on the TNSCUI dataset and DDTI dataset. Finally, for melanoma or nevus segmentation, we conducted experiments on the ISIC dataset [7, 27]. More details about the datasets can be found in the appendix.

3.2 Main Results

To verify the general performance of our proposed MSA model, we compare it with state-of-the-art (SOTA) segmentation methods on the multi-organ segmentation datasets AMOS and BTCV. The quantitative results are presented in Tables 1 and 2, respectively. In the tables, we compare MSA with the SAM model, as well as widely-used and well-recognized medical image segmentation methods, including nnUNet [19], TransUNet [3], UNetr [15], Swin-UNetr [14], EnsDiff [36], SegDiff [1], and MedSegDiff [37], all of which are based on either convolutional neural network (CNN) or transformer architectures. We also compare with fully fine-turned SAM on the training data, which is called MedSAM in [23]. We evaluate the segmentation performance using the Dice score. The detailed settings of the compared methods can be found in the appendix.

Table 1: The comparison of MSA with SOTA segmentation methods and original SAM over AMOS dataset evaluated by Dice Score. Best results are denoted as **bold**.

Methods	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	Panc.	RAG	LAG	Duo.	Blad.	Postc.	Avg
TransUNet	0.881	0.928	0.919	0.813	0.740	0.973	0.832	0.919	0.841	0.713	0.638	0.565	0.685	0.748	0.692	0.792
EnsDiff	0.905	0.918	0.904	0.732	0.723	0.947	0.838	0.915	0.838	0.704	0.677	0.618	0.715	0.673	0.680	0.786
SegDiff	0.885	0.872	0.891	0.703	0.654	0.852	0.702	0.874	0.819	0.715	0.654	0.632	0.697	0.652	0.695	0.753
UNetr	0.926	0.936	0.918	0.785	0.702	0.969	0.788	0.893	0.828	0.732	0.717	0.554	0.658	0.683	0.722	0.762
Swin-UNetr	0.959	0.960	0.949	0.894	0.827	**0.979**	0.899	0.944	0.899	0.828	0.791	0.745	0.817	0.875	0.841	0.880
nnUNet	0.965	0.959	0.951	0.889	0.820	0.980	0.890	0.948	0.901	0.821	0.785	0.739	0.806	0.869	0.839	0.878
MedSegDiff	0.963	**0.965**	0.953	0.917	0.846	0.971	0.906	0.952	0.918	0.854	0.803	0.751	**0.819**	0.868	0.855	0.889
SAM 1 point	0.632	0.759	0.770	0.616	0.382	0.577	0.508	0.720	0.621	0.317	0.085	0.196	0.339	0.542	0.453	0.493
SAM 3 points	0.733	0.784	0.786	0.683	0.448	0.658	0.577	0.758	0.625	0.343	0.129	0.240	0.325	0.631	0.493	0.542
SAM 10 points	0.857	0.855	0.857	0.800	0.643	0.811	0.749	0.842	0.677	0.538	0.405	0.516	0.480	0.789	0.637	0.699
MedSAM 1 point	0.671	0.803	0.825	0.687	0.541	0.712	0.671	0.785	0.703	0.607	0.531	0.588	0.729	0.814	0.833	0.700
MSA 1-point	**0.968**	0.961	**0.959**	**0.926**	**0.861**	0.971	**0.919**	**0.960**	**0.928**	**0.863**	**0.825**	**0.767**	0.803	**0.879**	**0.862**	**0.893**

Table 2: The comparison of MedSegDiff-V2 with SOTA segmentation methods over BTCV dataset evaluated by Dice Score. Best results are denoted as **bold**.

Model	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	Veins	Panc.	AG	Ave
TransUNet	0.952	0.927	0.929	0.662	0.757	0.969	0.889	0.920	0.833	0.791	0.775	0.637	0.838
EnsDiff	0.938	0.931	0.924	0.772	0.771	0.967	0.910	0.869	0.851	0.802	0.771	0.745	0.854
SegDiff	0.954	0.932	0.926	0.738	0.763	0.953	0.927	0.846	0.833	0.796	0.782	0.723	0.847
UNetr	0.968	0.924	0.941	0.750	0.766	0.971	0.913	0.890	0.847	0.788	0.767	0.741	0.856
Swin-UNetr	0.971	0.936	0.943	0.794	0.773	0.975	0.921	0.892	0.853	0.812	0.794	0.765	0.869
nnUNet	0.942	0.894	0.910	0.704	0.723	0.948	0.824	0.877	0.782	0.720	0.680	0.616	0.802
MedSegDiff	0.973	0.930	0.955	0.812	0.815	0.973	0.924	0.907	0.868	0.825	**0.788**	0.779	0.879
SAM 1 points	0.518	0.686	0.791	0.543	0.584	0.461	0.562	0.612	0.402	0.553	0.511	0.354	0.548
SAM 3 points	0.622	0.710	0.812	0.614	0.605	0.513	0.673	0.645	0.483	0.628	0.564	0.395	0.631
SAM 10 points	0.785	0.774	0.863	0.658	0.673	0.785	0.760	0.712	0.562	0.703	0.651	0.528	0.704
MedSAM 1 point	0.751	0.814	0.885	0.766	0.821	0.901	0.855	0.872	0.746	0.771	0.760	0.705	0.803
MSA 1 point	**0.978**	**0.935**	**0.966**	**0.823**	**0.818**	**0.981**	**0.931**	**0.915**	**0.877**	**0.811**	0.767	**0.809**	**0.883**

Table 1 and 2 show that the zero-shot performance of SAM is generally worse than that of fully-trained models on the target medical image segmentation tasks, regardless of the prompt given. Although this comparison may seem unfair (comparing SAM's zero-shot performance with fully-trained medical image models), SAM's zero-shot performance has been shown to outperform fully-trained models on nature image datasets. This suggests that SAM has worse zero-shot transfer capability on medical images, which has also been observed in many other studies [9, 16, 31].

Comparing the performance of SAM with different prompts, we observe that 3-point prompts perform slightly better than 1-point prompts, while 10-point prompts show a solid improvement over 3-point prompts. Although 10-point prompts still perform worse than the other fully-trained models, it is an impressive result considering the zero-shot setting. This highlights the strong potential of the SAM architecture for medical images. By exploiting this potential, MSA achieves a significant improvement over SAM given only a 1-point prompt. On the AMOS dataset, MSA achieves SOTA performance on 12 out of 15 organs and the best overall performance. On the BTCV dataset, MSA achieves SOTA performance on 11 out of 12 organs and the best overall performance. These results demonstrate that by using the correct fine-tuning technology and a super good pre-trained model, even on nature images, can be very beneficial to medical image segmentation, and can even lift its performance over specifically optimized models for medical image segmentation.
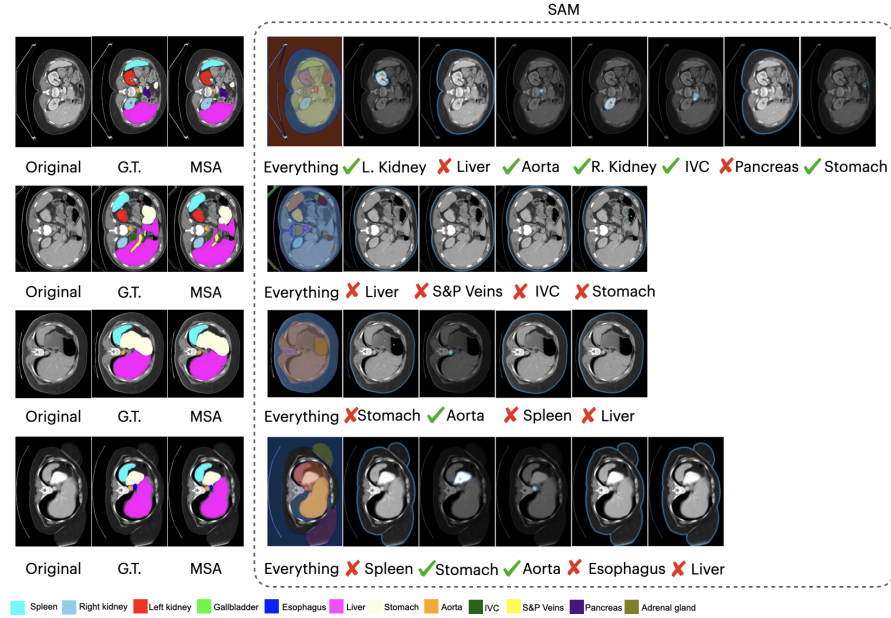

Fig. 2: Visual comparison of MSA and SAM on abdominal multi-organ segmentation. We use check mark to represent SAM correctly found the organ and cross to represent it lost.

Figure 2 presents a qualitative comparison of the performance of MSA and SAM. From the figure, it can be observed that MSA segments accurately on parts that are difficult to recognize by the human eye. On the other hand, SAM fails on many cases where the organ is actually quite clear. This once again demonstrates

the necessity of fine-tuning a general segmentation model on medical images to achieve optimal performance.

Table 3: The comparison of MSA with SAM and SOTA segmentation methods on different image modalities. The grey background denotes the methods are proposed for that/these particular tasks. Performance is omitted (-) if the algorithm fails over 70% of the samples.

	Optic-Cup		Brain-Turmor			Thyroid Nodule	
	Dice	IoU	Dice	IoU	HD95	Dice	IoU
ResUnet	80.1	72.3	78.4	71.3	18.71	78.3	70.7
BEAL	83.5	74.1	78.8	71.7	18.53	78.6	71.6
TransBTS	85.4	75.7	87.6	78.44	12.44	83.8	75.5
EnsemDiff	84.2	74.4	88.7	80.9	10.85	83.9	75.3
MTSeg	82.3	73.1	82.2	74.5	15.74	82.3	75.2
UltraUNet	83.1	73.78	84.5	76.3	14.03	84.5	76.2
SegDiff	82.5	71.9	85.7	77.0	14.31	81.9	74.8
nnUNet	84.9	75.1	88.5	80.6	11.20	84.2	76.2
TransUNet	85.6	75.9	86.6	79.0	13.74	83.5	75.1
UNetr	83.2	73.3	87.3	80.6	12.81	81.7	73.5
Swin-UNetr	84.3	74.5	88.4	81.8	11.36	83.5	74.8
MedsegDiff	85.9	76.2	**88.9**	**81.2**	**10.41**	84.8	76.4
SAM 1 points	-	-	63.2	58.6	25.53	-	-
SAM 3 points	-	-	65.5	61.7	24.87	-	-
MSA	**86.8**	**78.8**	87.6	81.2	12.46	**86.3**	**78.7**

We also compare MSA to state-of-the-art (SOTA) segmentation methods proposed for three specific tasks with different image modalities. The results are presented in Table 3. In the table, ResUnet [40] and BEAL [34] are proposed for optic cup segmentation, TransBTS [35] and EnsemDiff [36] are proposed for brain tumor segmentation, and MTSeg [13] and UltraUNet [6] are proposed for thyroid nodule segmentation. SegDiff, nnUNet, TransUNet, UNetr, Swin-UNetr and MedSegDiff are proposed for the general image segmentation.

The segmentation performance was evaluated using Dice score, IoU and HD95 metric. From the table, we can see MSA gains SOTA performance on Optic-Cup and Thyroid Nodule segmentation tasks, and outperforms most of the models on the brain-tumor segmentationn, showcasing its ability to generalize to various medical segmentation tasks and image modalities.

4   Conclusion

In this paper, we have extended the Segment Anything Model (SAM), a powerful general segmentation model, to medical image segmentation, and named it MSA. By employing parameter-efficient adaptation, a cost-effective fine-tuning technique, we have achieved significant improvement over the original SAM model and obtained state-of-the-art performance on 19 medical image segmentation tasks across 5 different image modalities. These results demonstrate the effectiveness of our approach to adaptation for medical images and the potential for transferring strong general segmentation models for medical applications.

We hope that this work can serve as a starting point towards advancing general medical image segmentation and inspire the development of new fine-tuning techniques.

References

1. Amit, T., Nachmani, E., Shaharbany, T., Wolf, L.: Segdiff: Image segmentation with diffusion probabilistic models. arXiv preprint arXiv:2112.00390 (2021)
2. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
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, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. arXiv preprint arXiv:2205.13535 (2022)
5. Chetoui, M., Akhloufi, M.A.: Explainable end-to-end deep learning for diabetic retinopathy detection across multiple datasets. Journal of Medical Imaging **7**(4), 044503–044503 (2020)
6. Chu, C., Zheng, J., Zhou, Y.: Ultrasonic thyroid nodule detection method based on u-net network. Computer Methods and Programs in Biomedicine **199**, 105906 (2021)
7. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 168–172. IEEE (2018)
8. Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al.: Bcn20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288 (2019)
9. Deng, R., Cui, C., Liu, Q., Yao, T., Remedios, L.W., Bao, S., Landman, B.A., Wheless, L.E., Coburn, L.A., Wilson, K.T., et al.: Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. arXiv preprint arXiv:2304.04155 (2023)
10. Fang, H., Li, F., Fu, H., Sun, X., Cao, X., Son, J., Yu, S., Zhang, M., Yuan, C., Bian, C., et al.: Refuge2 challenge: Treasure for multi-domain learning in glaucoma assessment. arXiv preprint arXiv:2202.08994 (2022)
11. Fang, X., Yan, P.: Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. IEEE Transactions on Medical Imaging **39**(11), 3619–3629 (2020)
12. Ghaffari, M., Sowmya, A., Oliver, R.: Automated brain tumor segmentation using multimodal brain scans: a survey based on models submitted to the brats 2012–2018 challenges. IEEE reviews in biomedical engineering **13**, 156–168 (2019)
13. Gong, H., Chen, G., Wang, R., Xie, X., Mao, M., Yu, Y., Chen, F., Li, G.: Multi-task learning for thyroid nodule segmentation with thyroid region prior. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 257–261. IEEE (2021)
14. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop. pp. 272–284. Springer (2022)

15. 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: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584 (2022)

16. He, S., Bao, R., Li, J., Grant, P.E., Ou, Y.: Accuracy of segment-anything model (sam) in medical image segmentation tasks. arXiv preprint arXiv:2304.09324 (2023)

17. He, X., Li, C., Zhang, P., Yang, J., Wang, X.E.: Parameter-efficient fine-tuning for vision transformers. arXiv preprint arXiv:2203.16329 (2022)

18. 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. arXiv preprint arXiv:2106.09685 (2021)

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. Nature methods **18**(2), 203–211 (2021)

20. Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023 (2022)

21. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)

22. Lin, Z., Zhang, Z., Chen, L.Z., Cheng, M.M., Lu, S.P.: Interactive image segmentation with first click attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13339–13348 (2020)

23. Ma, J., Wang, B.: Segment anything in medical images. arXiv preprint arXiv:2304.12306 (2023)

24. Mahadevan, S., Voigtlaender, P., Leibe, B.: Iteratively trained interactive segmentation. arXiv preprint arXiv:1805.04398 (2018)

25. Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., Yang, T., et al.: Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. Radiology: Artificial Intelligence **4**(5), e210315 (2022)

26. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)

27. Milton, M.A.A.: Automated skin lesion classification using ensemble of deep neural networks in isic 2018: Skin lesion analysis towards melanoma detection challenge. arXiv preprint arXiv:1901.10802 (2019)

28. Orlando, J.I., Fu, H., Breda, J.B., Van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., et al.: Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical image analysis **59**, 101570 (2020)

29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

30. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems **32** (2019)

31. Roy, S., Wald, T., Koehler, G., Rokuss, M.R., Disch, N., Holzschuh, J., Zimmerer, D., Maier-Hein, K.H.: Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model. arXiv preprint arXiv:2304.05396 (2023)
32. Tamkin, A., Liu, V., Lu, R., Fein, D., Schultz, C., Goodman, N.: Dabs: A domain-agnostic benchmark for self-supervised learning. arXiv preprint arXiv:2111.12062 (2021)
33. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data 5(1), 1–9 (2018)
34. Wang, S., Yu, L., Li, K., Yang, X., Fu, C.W., Heng, P.A.: Boundary and entropy-driven adversarial learning for fundus image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 102–110. Springer (2019)
35. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: Transbts: Multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 109–119. Springer (2021)
36. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.C.: Diffusion models for implicit image segmentation ensembles. arXiv preprint arXiv:2112.03145 (2021)
37. Wu, J., Fang, H., Zhang, Y., Yang, Y., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. arXiv preprint arXiv:2211.00611 (2022)
38. Xie, Y., Richmond, D.: Pre-training on grayscale imagenet improves medical image classification. In: Proceedings of the European conference on computer vision (ECCV) workshops. pp. 0–0 (2018)
39. Xing, Z., Wan, L., Fu, H., Yang, G., Zhu, L.: Diff-unet: A diffusion embedded network for volumetric segmentation. arXiv preprint arXiv:2303.10326 (2023)
40. Yu, S., Xiao, D., Frost, S., Kanagasingam, Y.: Robust optic disc and cup segmentation with deep learning for glaucoma detection. Computerized Medical Imaging and Graphics 74, 61–71 (2019)
41. Zaken, E.B., Ravfogel, S., Goldberg, Y.: Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. arXiv preprint arXiv:2106.10199 (2021)