Reproducibility Analysis: Reproduce the Top One Team Results

Li Zhi[1[0000−0002−8789−0151]], Yaqi Wang[2[0000−0002−4627−3392]], and Shuai Wang[1[0000−0003−3730−6401]]

[1] Hangzhou Dianzi University, Hangzhou, China
[2] Communication University of Zhejiang, Hangzhou, China
zhi.li@hdu.edu.cn
wangyaqi@cuz.edu.cn
shuaiwang.tai@gmail.com

Abstract. Many excellent solutions emerged in the competition. We chose to reproduce the Rank1 solution MedficientSAM, which uses the EfficientViT model to replace the heavy image encoder in SAM and then extracts knowledge from the MedSAM model on the challenge training set. The test results show that we successfully reproduced the Top One Team solution. During the verification process, due to the limitations of our hardware equipment, we reduced the batch size to 1/4 of the original solution while training. Besides, Additional X-Ray validation images were added in the post-challenge phase, resulting in a decline in the model performance. The average DSC and NSD scores of our reproduced scheme on the public validation set are 0.8516 and 0.8668 respectively, slightly lower than the average DSC and NSD scores of the original scheme of 0.8642 and 0.8795. However, we still achieved much better results than the baseline average DSC score of 83.23 and NSD score of 82.71. It also proves the reproducibility of the top One team solution. Our detailed experiment logs, trained weights, and docker are publicly available at: https://github.com/RicoLeehdu/medficientsam-reproduce.

Keywords: Medical image segmentation · Edge AI

1 Introduction

The “Segment Anything In Medical Images On Laptop” challenge aims to develop universal promptable medical image segmentation models that can be deployed on laptops or edge devices without relying on GPUs. Specifically, participants are tasked with creating a lightweight, bounding box-based segmentation model. The challenge also introduces a baseline model, LiteMedSAM, which replaces the heavy image encoder in MedSAM with TinyViT [7], a scaled-down vision transformer model using a progressive contraction approach [3]. The challenge provides a large training dataset with over one million image-mask pairs, covering 11 types of medical images, along with more than 20 types of cancer. Many works have introduced lighter models to address computational constraints by replacing the heavy image encoder of SAM. In natural image processing, notable examples include MobileSAM [9] and EfficientViT-SAM [10]. MobileSAM utilizes TinyViT as a lightweight image encoder, similar to LiteMedSAM. EfficientViT-SAM, on the other hand, replaces traditional softmax attention [6] with lightweight ReLU linear attention [3], reducing computational complexity from quadratic to linear while maintaining functionality. The benchmarks in [10] indicate that EfficientViT-SAM offers higher throughput than MobileSAM, despite having more parameters, and also delivers superior segmentation accuracy, even outperforming the original SAM.

2 Method

2.1 Preprocessing

We chose to replicate the Rank1 solution MedficientSAM, which uses the EfficientViT model to replace the heavy image encoder in SAM and then extracts knowledge from the MedSAM model on the challenge training set.

2.2 Post-processing

The binary masks output by MedficientSAM have a fixed size of 256 × 256. We first resize these output masks to match the input size of the image encoder, then crop out the padded zeros, and finally resize them back to their original resolution.

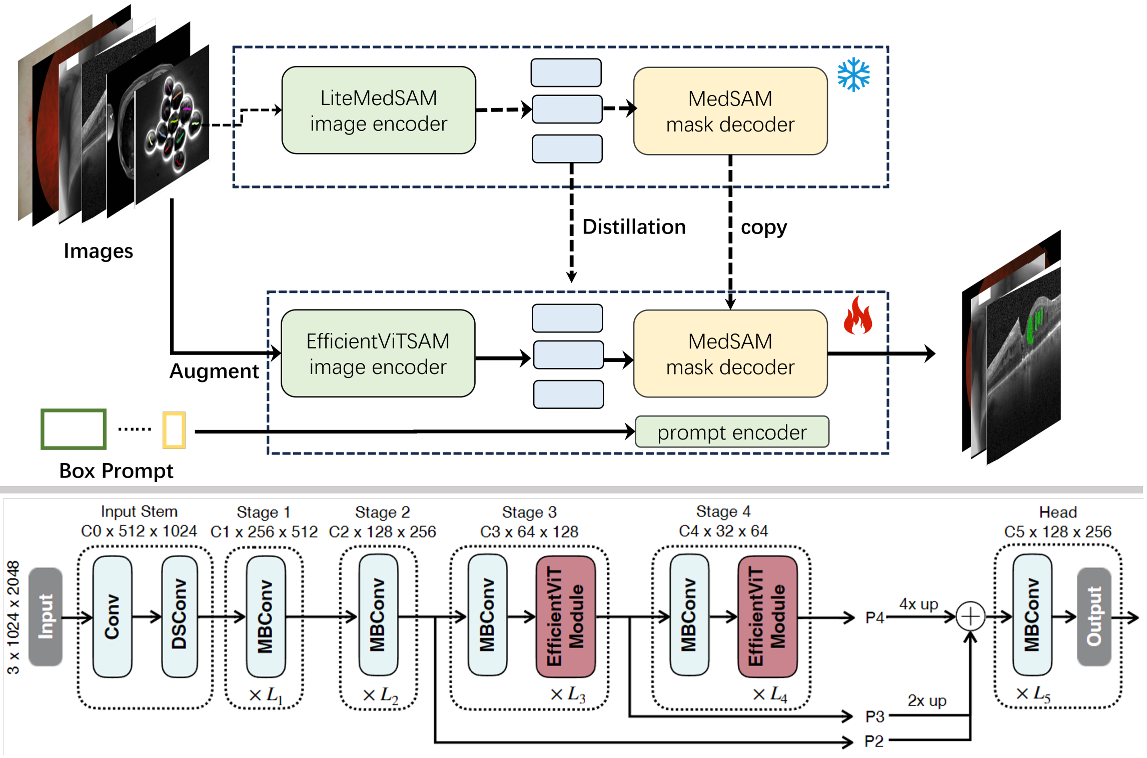

Fig. 1. MedficientSAM training pipline (top) and EfficientViT-SAM-L1's macro architecture (bottom).Top: The training pipeline contains two stages the distillation stage and the fine-tuning stage. Bottom: "ResBlock" refers to the basic building block from ResNet34 [2]. "FMBConv" refers to the fused MBConv block from [5]. "EfficientViT Module" is the building block from [1].

3 Experiments

3.1 Dataset and evaluation measures

The evaluation metrics include two accuracy measures: Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD), alongside running time as an efficiency measure. These metrics collectively contribute to the ranking computation. The evaluation platform is CPU-only to simulate edge devices, running on an Intel(R) Xeon(R) Silver 4114 CPU at 2.20GHz with 10 cores. Furthermore, the memory usage is constrained to a maximum of 8 GB.

3.2 Implementation details

Environment settings The development environments and requirements are presented in Table 1.

Table 1. Development environments and requirements.

System	Ubuntu 20.04.6 LTS
CPU	Intel(R) Xeon(R) Silver 4114
RAM	128GB
GPU (number and type)	One NVIDIA 4080 16G
CUDA version	12.1
Programming language	Python 3.10
Deep learning framework	torch 2.2.2, torchvision 0.17.2
Code	https://github.com/RicoLeehdu/ExpertsSAM/tree/master

Training protocols We adopted the same solution as the Top One team, except that due to hardware equipment limitations, we adopted 1/4 of the batch size of the original solution, which resulted in longer training time and a slight decrease in effect. The training protocols from the distillation and fine-tuning stage are listed in Table 2 and Table 3.

4 Reproducible Results and Discussion

The reproduction results from Table 4 and Table 5 shows that the average DSC and NSD scores of our reproduced scheme on the public validation set are 0.8516 and 0.8668 respectively, slightly lower than the average DSC and NSD scores of the original scheme of 0.8642 and 0.8795 respectively. Besides, additional X-Ray validation images were added in the post-challenge phase, resulting in a decline in the model performance. Due to the limitations of our hardware equipment, we reduced the batch size to 1/4 of the original solution during the distillation and fine-tuning stages, which resulted in a decline in the model training performance. However, we still achieved much better results than the baseline average DSC score of 83.23 and NSD score of 82.71. Score better results It also proves the reproducibility of the Top one team solution.

Table 2. Training protocols for distillation stage.

Teacher Model	MedSAM [4]
Student Model	EfficientViT-L1[15]
Data augmentation	Horizontal Flipping and Vertical Flipping
Batch size	2
Patch size	512×512×3
number works	32
Total epochs	8
Optimizer	AdamW with weight decay set to 0.0005
Initial learning rate (lr)	0.075
Lr decay schedule	ReduceLROnPlateau
Training time	91 hours
Lr decay schedule	decay the Lr by 0.5 every epoch
Loss function	L2
Number of model parameters	43.59M
Number of model flops	49.23G

Table 3. Training protocols for fine-tuning stage.

Model	MedficientSAM-L1
Data augmentation	Horizontal Flipping, Vertical Flipping, and Shift Scale Rotate
Patch size	$512 \times 512 \times 3$
Batch size	8
Total epochs	8
Optimizer	AdamW [13] with default settings
Initial learning rate (lr)	2×10^{-6}
Lr decay schedule	Cosine Annealing [12]
Training time	104 hours
Number of model parameters	47.65M
Number of flops	51.05G

Table 4. Quantitative evaluation results. Those with the -R suffix are reproducible results, the optimal results are given in red, and the suboptimal results are given in blue.

Target	LiteMedSAM		Distillation		Distillation-R		No Augmentation		No Augmentation-R		MedficientSAM-L1		MedficientSAM-L1-R	
	DSC(%)	NSD(%)	DSC(%)	NSD(%)	DSC(%)	NSD(%)	DSC(%)	NSD(%)	DSC(%)	NSD(%)	DSC(%)	NSD(%)	DSC(%)	NSD(%)
CT	92.26	94.90	91.13	93.75	92.15	94.74	92.24	94.71	92.69	95.50	92.15	94.80	93.19	95.78
MR	89.63	93.37	85.73	89.75	87.87	91.40	87.25	90.88	88.54	92.21	86.98	90.77	89.51	92.99
PET	51.58	25.17	70.49	54.52	68.30	50.17	72.05	56.26	61.06	49.13	73.00	58.03	66.97	52.52
US	94.77	96.81	84.43	89.29	84.52	89.37	81.99	86.74	82.41	87.16	82.50	87.24	81.39	86.09
X-Ray	75.83	80.39	78.92	84.64	75.40	80.38	79.88	85.73	78.04	83.10	80.47	86.23	75.78	80.88
Dermoscopy	92.47	93.85	92.84	94.16	92.54	93.88	94.24	95.62	93.71	95.19	94.16	95.54	93.17	94.62
Endoscopy	96.04	98.11	96.88	98.81	95.92	98.16	96.05	98.33	95.58	98.07	96.10	98.37	94.62	97.26
Fundus	94.81	96.41	94.10	95.83	93.85	95.54	94.16	95.89	94.27	96.00	94.32	96.05	94.16	95.90
Microscopy	61.63	65.38	75.63	82.15	75.90	82.45	78.76	85.22	78.09	84.48	78.09	84.47	77.67	84.11
Average	83.23	82.71	85.57	86.99	85.16	86.23	86.29	87.71	84.93	86.76	86.42	87.95	85.16	86.68

Table 5. Segmentation efficiency results on the public validation set. The computational metrics from MedficientsSAM are obtained on an Intel(R) Core(TM) i9-10900K and from our reproduced results are obtained on an Intel(R) Xeon(R) Silver 4114, except for MedSAM, which can not run on CPU. Those with the -R suffix are reproducible results.

Method	Res.	#Params	#FLOPs	DSC	NSD	DSC-R	NSD-R	2D Runtime	3D Runtime	2D Memory Usage	3D Memory Usage
MedSAM	1024	93.74M	488.24G	84.91	86.46	84.91	86.46	N/A	N/A	N/A	N/A
LiteMedSAM	256	9.79M	39.98G	83.23	82.71	83.23	82.71	5.1	42.6	1135	1241
MedficientSAM-L0	512	34.79M	36.80G	85.85	87.05	84.93	86.76	0.9	7.4	448	687
MedficientSAM-L1	512	47.65M	51.05G	86.42	87.95	85.16	86.68	1.0	9.0	553	793
MedficientSAM-L2	512	61.33M	70.71G	86.08	87.53	85.07	86.63	1.1	11.1	663	903

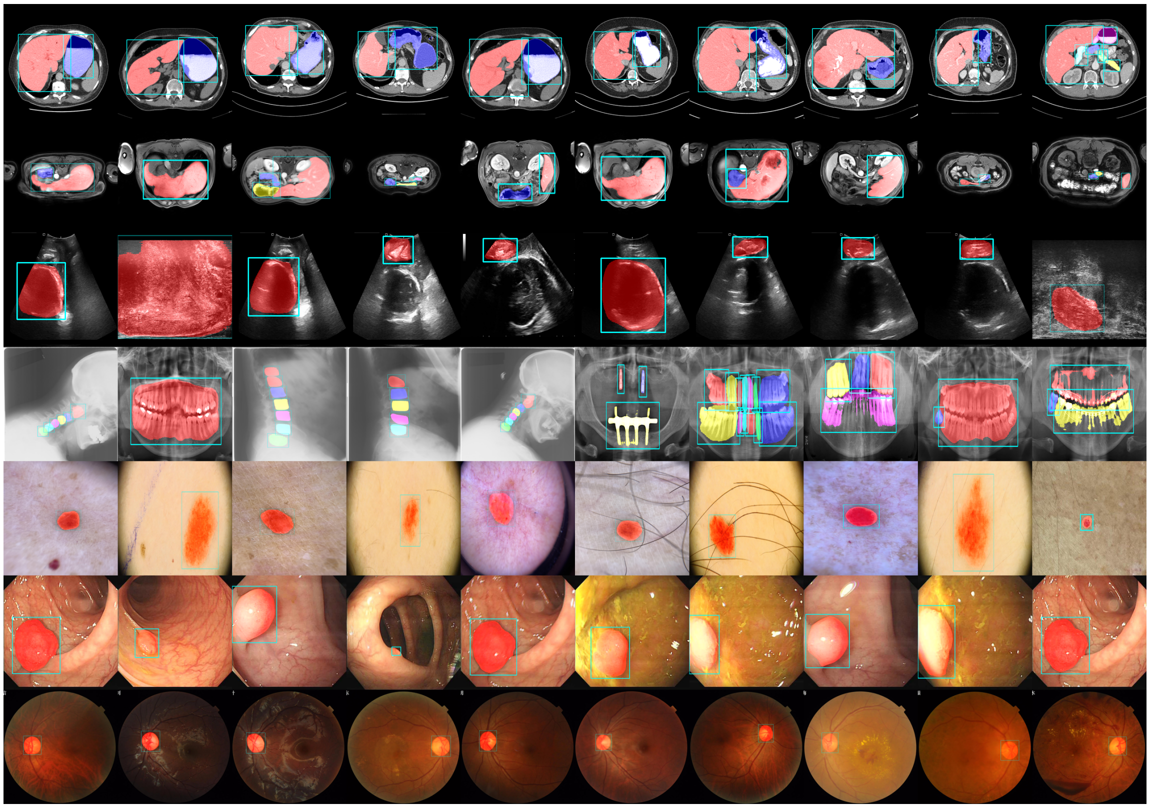

Fig. 2. The good and bad cases in fundus segmentation tasks.

Table 6. Performance of different modalities in test results.

Modality	CT	MR	X-Ray	Endoscopy	Fundus	Microscope	OCT	PET	US
DSC	69.25	69.77	80.89	95.01	88.17	88.11	81.87	72.31	89.50
NSD	75.80	69.14	91.04	97.47	90.29	89.80	87.99	63.08	93.82

Table 7. Runtime for different modalities in Seconds

Modality	CT	MR	X-Ray	Endoscopy	Fundus	Microscope	OCT	PET	US
RunTime	9.56	4.62	1.73	1.39	1.51	1.88	1.46	3.43	1.89

4.1 Quantitative results on validation set

The reproduction results from Table 4 and Table 5 show that in most cases, the reproduced DSC and NSD values are very close to the original results, and in some cases improved. For example, in the DSC and NSD reproduction results of CT, and Fundus, the reproduced DSC values and NSD values sometimes exceed the original results and achieve higher accuracy. Modalities with excellent performance In cytomicroscopy (Microscopy), chest X-ray (X-Ray), and abdominal endoscopy (Endoscopy), MedficientSAM performs very well, with DSC values between 94% and 98%, which is above average. This may be because these images have high resolution, clear borders, and large target areas. In addition, RGB images are easier to segment than grayscale images due to better color discrimination. Challenging modalities In some challenging modalities, such as PET, the model's performance is much lower than The average level is only 64% to 68%. The poor performance of the model on these images may be due to the characteristics of the imaging modality, such as PET having a different color scale than other types. In addition, low resolution will also make segmentation less effective because the image will become blurry and the boundaries will not be clear after resizing. Conclusion The reproduction results show that MedficientSAM can reach or exceed the original results in some modalities, but still needs improvement in some challenging modalities.

4.2 Qualitative results on validation set

Fig. 2 illustrates the reproducible segmentation results. It can be observed that the segmentation results of our reproduced model are reliable in most modalities, but the segmentation effect of curved tomography of teeth in X-Ray images is not good. This may be because the model has not been trained on similar data modalities and cannot segment structured teeth from curved tomography.

4.3 Results on final testing set

The results on the final testing set are listed in Table 6 and runtime of each modalities is listed in Table 7. The results show that we successfully reproduced the Top One Team's solution.

5 Conclusion

Due to the limitations of our hardware equipment, we reduced the batch size to 1/4 of the original solution during the distillation and fine-tuning stages, which resulted in a decline in model training effect, but we still achieves better results than the baseline average DSC score of 83.23 and NSD score of 82.71. Score better results It also proves the reproducibility of the Top one team solution. The average inference time is 1.0083 seconds for 2D images and 8.9585 seconds for 3D images. Our detailed experiment logs are publicly available at:

Acknowledgements We thank all the data owners for making the medical images publicly available and CodaLab [8] for hosting the challenge platform.

References

1. Cai, H., Li, J., Hu, M., Gan, C., Han, S.: Efficientvit: Lightweight multi-scale attention for on-device semantic segmentation. arXiv preprint arXiv:2205.14756 (2023) 2
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 2
3. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are rnns: Fast autoregressive transformers with linear attention. In: International conference on machine learning. pp. 5156–5165. PMLR (2020) 1
4. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024) 4
5. Tan, M., Le, Q.: Efficientnetv2: Smaller models and faster training. In: International conference on machine learning. pp. 10096–10106. PMLR (2021) 2
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 1
7. Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., Yuan, L.: Tinyvit: Fast pretraining distillation for small vision transformers. In: European Conference on Computer Vision. pp. 68–85. Springer (2022) 1
8. Xu, Z., Escalera, S., Pavão, A., Richard, M., Tu, W.W., Yao, Q., Zhao, H., Guyon, I.: Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. Patterns **3**(7), 100543 (2022) 6
9. Zhang, C., Han, D., Qiao, Y., Kim, J.U., Bae, S.H., Lee, S., Hong, C.S.: Faster segment anything: Towards lightweight sam for mobile applications. arXiv preprint arXiv:2306.14289 (2023) 1
10. Zhang, Z., Cai, H., Han, S.: Efficientvit-sam: Accelerated segment anything model without performance loss. In: CVPR Workshop: Efficient Large Vision Models (2024) 1