This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

LGA: A Language Guide Adapter for Advancing the SAM Model's Capabilities in Medical Image Segmentation

Jihong Hu[1], Yinhao Li[2], Hao Sun[3], Yu Song[2], Chujie Zhang[1], Lanfen Lin[3](✉), and Yen-Wei Chen[2](✉)

[1] Graduate School of Information Science and Engineering, Ritsumeikan University, Japan
{gr0609ik, gr0696fh}@ed.ritsumei.ac.jp
[2] College of Information Science and Engineering, Ritsumeikan University, Japan
{yin-li@fc, yusong@fc, chen@is}.ritsumei.ac.jp,
[3] College of Computer Science and Technology, Zhejiang University, Hangzhou, China
{sunhaoxx,llf}@zju.edu.cn

Abstract. In addressing the unique challenges of medical image segmentation, foundation models like the Segment Anything Model (SAM), originally developed for natural image, often falter due to the distinct nature of medical images. This study introduces the Language Guide Adapter (LGA), a paremeter efficient fine-tuning approach that extends SAM's utility to medical segmentation tasks. Through the integration of textual data from medical reports via a pretrained Bert model into embeddings, LGA combines these embeddings with the image features in SAM's image encoder using Feature Fusion Modules (FFM). Our method significantly enhances model performance and reduces computational overhead by freezing most parameters during the fine-tuning process. Evaluated on the CT-based MosMedData+ and the X-ray dataset QaTa-COV19, LGA demonstrates its effectiveness and adaptability, achieving competitive results with a significant reduction in the number of parameters required for fine-tuning compared to SOTA medical segmentation models. This enhancement underscores the potential of foundation models, leveraging the integration of multimodal knowledge as a pivotal approach for application in specialized medical tasks, thus charting a course towards more precise and adaptable diagnostic methodologies. The code is avaliable at https://github.com/JiHooooo/LGA.

Keywords: Medical image segmentation · Foundation model · Vision-language model · Parameter efficient fine-tune.

1 Introduction

The segmentation of medical images represents a pivotal task in computer-assisted diagnosis. In recent years, deep learning models have achieved notable success in this field, enabling fully automated segmentation of objects of interest[22,15,4,27].

Nonetheless, these models are heavily dependent on large, annotated datasets for training—a requirement that presents substantial challenges in the medical field. The need for expert annotations from professional doctor renders the data preparation phase both time-consuming and expensive. Additionally, the diversity of medical imaging modalities, along with variations in imaging equipment and operational methods across different hospitals, significantly challenges the generalizability of these models.

Transitioning from these traditional approaches, the recent advent of foundation models marks a paradigm shift towards overcoming such limitations. Initially excelling in Natural Language Processing (NLP), models like GPT[20] and BERT[8] have showcased their prowess in text understanding and generation. By leveraging zero-shot learning and prompt engineering, these foundation models perform well across various tasks with minimal data, demonstrating their adaptability and efficiency without extensive training. In the field of image segmentation, SAM[16] emerges as the pioneering foundation model. Equipped with an expansive dataset of 11 million labeled images and an interactive segmentation framework, SAM achieves remarkable zero-shot capabilities on natural images. However, studies[14,10] have indicated SAM's underperformance in the medical imaging domain, highlighting the need for further improvement to bridge the applicability gap.

Given the substantial parameter volume inherent in foundation models, comprehensively fine-tuning such models requires extensive computational resources, and when attempted with limited datasets, often results in suboptimal performance. Consequently, there's a growing focus on parameter-efficient fine-tuning strategies. For instance, MedSAM[19] adopts a strategy where it freezes the parameter-heavy image encoder within SAM, opting to fine-tune only the mask decoder using medical data. Meanwhile, [24,12,5] leverage Adapter technology, inserting lightweight, learnable modules into the existing model for fine-tuning the image encoder. Techniques like Low-Rank Adaptation (LoRA) are also used to adjust the encoder [25,26]. However, these methodologies remain primarily confined to the imaging modality.

Contrasting with the limitations inherent to the imaging modality alone, leveraging the intrinsic value of medical reports presents a unique opportunity in medical image analysis. Since each patient's medical imaging is accompanied by a corresponding medical report, acquiring these reports incurs no additional cost, unlike the augmentation of segmentation annotations. Furthermore, the imaging quality of medical images is generally lower than that of natural images, with boundaries between different regions appearing more blurred. However, medical reports, enriched with expert knowledge, compensate for the deficiencies in image quality. [13] developes an attention-based framework for learning both global and local representations by contrasting image sub-regions with words in the paired report. [23] improves polyp segmentation by incorporating text-guided features like polyp size and count. [17] leverages Bert to extract textual features of medical reports and employs a hybrid network of CNNs and Transformers to

fuse textual and visual features. These methods still require designing complex networks from scratch to fuse textual and image information.

In contrast, we propose a more flexible multimodal model framework that retains the parameters of foundation models like SAM and BERT used in our experiments. Our approach uses lightweight feature interaction modules for feature fusion. For each new task, only these modules and the task-specific prediction head need fine-tuning. This method enhances model performance while reducing computational and storage resources. Our contributions are summarized in three key areas:

1. We introduce a parameter-efficient fine-tuning method for SAM foundational models, called language guide adapter(LGA), which incorporates textual information into the fine-tuning process, significantly enhancing the model's performance on specific medical segmentation tasks. To the best of our knowledge, this is the first study to implement adapter technology for incorporating textual information into SAM foundation model.
2. We have developed an efficient feature fusion module(FFM) that combines cross-attention mechanisms with Multi Layer Perceptron(MLP) networks to achieve the integration of textual and visual information.
3. Through comprehensive experiments on the X-ray QaTa-COV19 and CT MosMedData+ datasets, our study showcases it's exceptional adaptability across various medical imaging modalities. It not only surpasses SOTA algorithms in terms of performance but also achieves this with a significantly reduced number of training parameters.

2 Methodology

The SAM model comprises three primary components: an image encoder with a vision transformer architecture for feature extraction[9] , a prompt encoder for encoding prompt information, and a mask decoder for generating final predictions using both prompt and image features. The majority of the SAM model's parameters are concentrated in the image encoder. As illustrated in 1, we remove the prompt encoder, tailoring the model for automated end-to-end segmentation tasks. Textual features are extracted from medical reports using the pretrained Bert model and then combined with image features via the LGA, which integrates multiple FFMs. Moreover, during the fine-tuning phase, we specifically adjust only the parameters of the LGA and the mask decoder. By freezing the entire image encoder, we significantly reduce the number of parameters that need fine-tuning.

2.1 LGA framework

The LGA comprises N FFMs. For each module, designated as F^n_{FFM} for the n-th module, a textual feature L_n and an image feature V_m are inputted, yielding an updated textual feature $\hat{L}_n$ and image feature $\hat{V}_m$, $(\hat{L}_n, \hat{V}_m) = F^n_{FFM}(L_n, V_m)$. Here, V_m represents the feature output from the m-th transformer block within

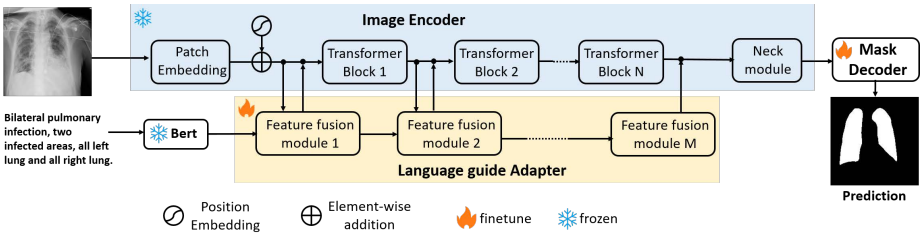

Fig. 1. LGA Framework Overview: Extracting image features via image encoder and text features via Bert, with LGA fusing both for prediction by the mask decoder. During fine-tuning, only LGA and mask decoder are adjusted, freezing Bert and image encoder.

the image encoder, and $\hat{V}_m$ is subsequently fed back into the image encoder for further processing. The updated textual feature $\hat{L}_n$ is then utilized as the input for $F_{FFM}^{(n+1)}$, facilitating continuous feature fusion across the sequence of FFMs. Notably, the Bert model initially processes the medical reports into a summarized feature L_0 by converting and averaging word features, as detailed in Eq.1. Z and F represent the length of the medical report T and the length of Bert feature respectively.

$$L_n = \begin{cases} \sum_{j=1}^{Z} Bert(T) \in R^{1\times F}, & n = 0 \\ \hat{L}_{n-1}, & n > 0 \end{cases} \tag{1}$$

2.2 FFM structure

The FFM, depicted in Fig.2, stands as a crucial component within the LGA framework. It employs a dual cross-attention structure integrated with two MLPs to facilitate the fusion process. The entire computation process is illustrated in Eq.2, Eq.3, Eq.4 and Eq.5, with $norm(\cdot)$ denoting Layer Normalization[2]. The learnable parameter γ^n is initially configured to 0. This setup aims to maintain the pre-trained image encoder's feature extraction integrity in the training's early stages[6].

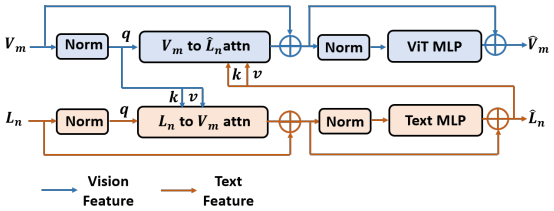

Fig. 2. The structure of feature fusion module.

$$L'_n = L_n + CrossAtten(norm(L_n), norm(V_m)) \tag{2}$$

$$\hat{L}_n = L'_n + MLP_{text}(norm(L'_n)) \tag{3}$$

$$V'_m = V_m + \gamma^n CrossAtten(norm(V_m), norm(\hat{L}_n)) \tag{4}$$

$$\hat{V}_m = V'_m + MLP_{vit}(norm(V'_m)) \tag{5}$$

2.3 Training process

For fine-tuning, we use a dataset $\{(X^d, Y^d, T^d)\}_{d=1}^{D}$ comprising images X^d, labels Y^d, and related medical reports T^d, with D samples in total. The Bert model transforms medical reports into text feature L_0^d. Then, the image encoder, integrating the LGA, extracts features from inputs X^d and L_0^d, which are then processed by the Mask decoder to yield the segmentation predictions P^d. We calculate loss using a mix of Dice and cross-entropy loss, detailed in Eq.6 and Eq.7, where K and C are pixel and class counts. P_{kc} is the probability of pixel k in class c, and Y_{kc} represents whether pixel k belongs to category c. The total loss formula is $l_{total} = 0.5l_{Dice} + 0.5l_{CE}$.

$$l_{Dice} = 1 - \sum_{k=1}^{K}\sum_{c=1}^{C} \frac{1}{KC} \frac{2\,|P_{kc} \cap Y_{kc}|}{(|P_{kc}| + |Y_{kc}|)} \tag{6}$$

$$l_{CE} = -\sum_{k=1}^{K}\sum_{c=1}^{C} \frac{1}{K} \cdot Y_{kc} log(P_{kc}) \tag{7}$$

3 Experiments

3.1 Datasets

We evaluated our model using two datasets: MosMedData+[1,11] and QaTa-COV19[7]. MosMedData+ consists of 2729 lung CT slices with COVID-19 findings, annotated with binary masks highlighting regions such as ground-glass opacifications. The QaTa-COV19 dataset contains 9258 chest X-rays annotated for COVID-19 lesions. Text annotations detailing aspects like infection presence, number of affected regions, and their specific locations (e.g., "Bilateral pulmonary infection, two infected areas, upper left lung and upper right lung") were added by [17]. For dataset partitioning, both MosMedData+ and QaTa-COV19 datasets were divided into training, validation, and test sets with 2,183, 273, 273, and 5,716, 1,429, 2,113 samples respectively, aligning with [17].

3.2 Implementations

In terms of data augmentation, we employed random cropping, rotation, translation, scaling, and adjustments to brightness and contrast. The specific parameters are detailed in the appendix. For model training, the batch size was set to 2, with AdamW[18] as the optimizer. The initial learning rate was set to 1e-4, with

a weight decay of 0.001. Regarding learning rate adjustment, we applied a linear warm-up for the first 1,000 iterations, followed by a cosine decay of the learning rate thereafter. Training was conducted over 50 epochs, utilizing a single RTX 4090 graphics card.

In our model structure, we utilized the ViT-B architecture, initializing it with pretrained SAM parameters from the SA-1B dataset[16]. The LGA includes four FFMs: the first three interact with the input features of the 1st, 5th, and 9th transformer blocks of ViT-B, respectively, while the final FFM engages with the output features of ViT-B's last transformer block. Within this last FFM, we omitted the cross-attention mechanism and the MLP, both initially intended to enhance text features. The detailed parameters for the FFMs are provided in the appendix.

3.3 Quantitative results

Table 1. Quantitative comparison of our proposed method with other SOTA medical image segmentation results

Method	Text	Param(M)	QaTa-COV19		MosMedData+	
			Dice(%)	mIoU(%)	Dice(%)	mIoU(%)
U-Net[22]	✗	14.8	79.02	69.46	64.60	50.73
UNet++[27]	✗	74.5	79.62	70.25	71.75	58.39
nnUNet[15]	✗	19.1	80.42	70.81	72.59	60.36
TransUNet[4]	✗	105	78.63	69.13	71.24	58.44
Swin-Unet[3]	✗	82.3	78.07	68.34	63.29	50.19
MedSAM[5]	✗	4.06	73.05	61.97	50.91	37.13
TGANet[23]	✓	19.8	79.72	70.58	72.06	59.73
CLIP[21]	✓	87.0	79.81	70.66	71.97	59.64
GLoRIA[13]	✓	45.6	79.94	70.68	72.42	60.18
LViT-T[17]	✓	29.7	83.66	75.11	74.57	61.33
LGA(our)	✓	8.24	**84.65**	**76.23**	**75.63**	**62.52**

In our quantitative analysis, the LGA model was evaluated against SOTA medical segmentation models using Dice and mIoU metrics. This comparison spanned text-integrated models (TGANet, CLIP, GLoRIA, LViT-T), image-only models (UNet, UNet++, nnUNet, TransUNet, Swin-Unet), and MedSAM, which fine-tunes only the SAM mask decoder. Results, shown in Table.1, indicate LGA's superior performance and efficiency: it achieved the highest scores in both metrics on all datasets and required significantly fewer training parameters than all models except MedSAM. Specifically, LGA improved Dice and mIoU on QaTa-COV19 by 11.60% and 14.26%, and on MosMedData+ by 24.72% and 25.39%, respectively, compared to MedSAM. Against the most competitive model, LViT-T, LGA recorded improvements of 0.99% in Dice and 1.12% in

mIoU on QaTa-COV19, and 1.06% in Dice and 1.19% in mIoU on MosMedData+, with a 72.3% reduction in fine-tuned parameters.

3.4 Model results visualization

We also visualized the model's results, as shown in Fig. 3. From left to right, the sequence is textual information, original image, label, and the results from the LGA, MedSAM, LViT, and nnUNet models. It is evident that, due to the lack of textual information, nnUNet and MedSAM are prone to inaccuracies in determining the position and number of infected regions. Compared to LViT, which also integrates textual information, the segmentation results of our proposed LGA are more precise.

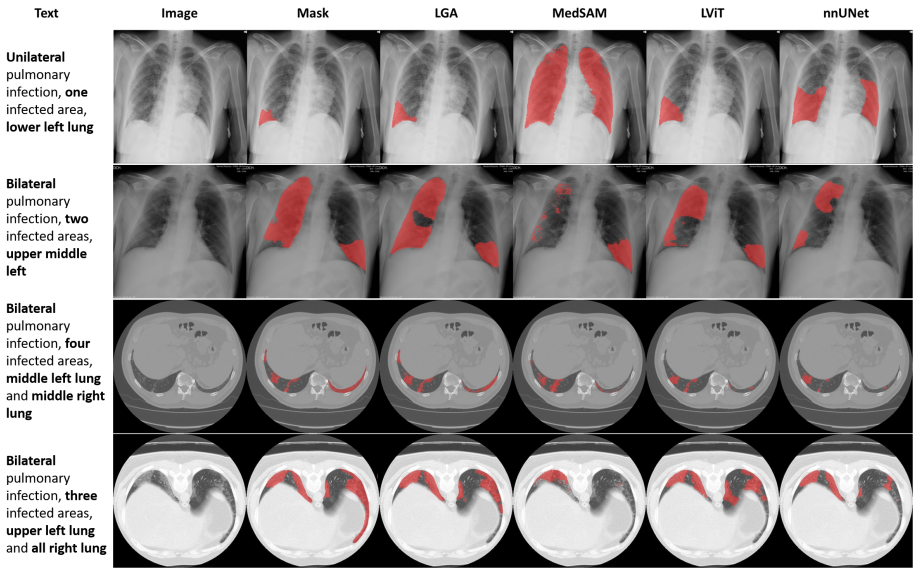

Fig. 3. Qualitative comparison of our proposed LGA with MedSAM, LViT and nnUNet

3.5 Ablation study

In the ablation study performed on the QaTa-COV19 dataset, our baseline involved solely fine-tuning the mask decoder. We first assessed the performance impact of integrating the LGA fusion module and textual information. Following this, we investigated the influence of both the structure and number of FFMs within the LGA on the model's effectiveness.

In Table.2, we progressively enhance the baseline by first adding textual information (following the original SAM's approach of directly inputting Bert-converted text into the Mask decoder), then by solely introducing the LGA without text (replacing BERT's text features with learnable query), and finally

by fully implementing the LGA strategy. Detailed test conditions are in the appendix. The results show that both LGA's improvement of the Image Encoder's feature extraction and the addition of textual information positively affect performance, with the combination of both achieving the best outcomes.

For analysing the FFM structure's impact, we sequentially added Dual cross attention and separate MLPs for image and text features on top of the baseline. We found that each component contributes to the final performance, as shown in Table.3.

Table 2. The ablation study on the QaTa-COV19 Dataset

LGA	Text	Param(M)	Dice(%)	mIoU(%)
		4.06	73.05	61.97
	✓	4.20	78.62	68.10
✓		8.24	80.37	70.83
✓	✓	8.24	**84.65**	**76.23**

Table 3. The influence of FFM structure on the QaTa-COV19 Dataset

Dual Cross	ViT MLP	Text MLP	Param (M)	Dice (%)	mIoU (%)
			4.06	73.05	61.97
✓			6.15	81.65	72.01
✓	✓		7.34	84.09	75.47
✓	✓	✓	8.24	**84.65**	**76.23**

Table 4. The influence of FFM number on the QaTa-COV19 Dataset

Number	Param(M)	Dice(%)	mIoU(%)
0	4.06	73.05	61.97
1	4.66	81.21	71.51
2	5.85	84.23	75.47
3	7.04	84.43	75.85
4	8.24	84.65	**76.23**
5	9.43	**84.68**	76.08

In Table.4, we assessed how varying the number of FFMs in the LGA affects performance with FFM placement specifics in the appendix. Compared to the baseline that only adjusts the mask decoder, integrating one FFM into the image encoder substantially enhanced performance, raising Dice by 8.16% and mIoU by 9.54%, with a minimal parameter increase of 0.60M. Additional FFMs further improved results, but the benefit plateaued after four FFMs, suggesting a saturation point. Thus, for efficiency, subsequent experiments used four FFMs.

4 Conclusion

In this paper, addressing the performance degradation of SAM foundation models in medical segmentation tasks, we introduce a parameter-efficient fine-tuning strategy named LGA, which boosts model performance by incorporating textual information. This approach also serves as an efficient paradigm for integrating textual and visual modalities, utilizing adapter technology to fuse features from both vision and language foundation models for specific tasks. We validated our method on two medical imaging datasets from different modalities:

the CT dataset MosMedData+ and the X-ray dataset QaTa-COV19, demonstrating its applicability across diverse medical imaging data types. Compared to other SOTA medical segmentation models that focus solely on the image modality or combine image and textual modalities, our LGA method not only achieves SOTA performance but also significantly reduces the amount of parameters required for fine-tuning. However, our current approach relies on textual input during inference, which can be a limitation when such information is not available. We plan to address this issue in future work.

Acknowledgments. This work was supported in part by the Grant in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports (MEXT) under the Grant Nos. 20KK0234, 21H03470, and 20K21821, and in part by Zhejiang Provincial Natural Science Foundation of China (No. LZ22F020012).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Andreychenko, A., Pavlov, N., Vladzymyrskyy, A., Ledikhova, N., Gombolevskiy, V., Gelezhe, P., Gonchar, A., Chernina, V.Y.: Mosmeddata: Chest ct scans with covid-19 related findings dataset. arXiv preprint arXiv:2005.06465 (2020)
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
3. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
4. 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)
5. Chen, T., Zhu, L., Ding, C., Cao, R., Zhang, S., Wang, Y., Li, Z., Sun, L., Mao, P., Zang, Y.: Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more. arXiv preprint arXiv:2304.09148 (2023)
6. Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., Qiao, Y.: Vision transformer adapter for dense predictions. In: The Eleventh International Conference on Learning Representations (2022)
7. Degerli, A., Kiranyaz, S., Chowdhury, M.E., Gabbouj, M.: Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 2306–2310. IEEE (2022)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

10. 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)
11. Hofmanninger, J., Prayer, F., Pan, J., Röhrich, S., Prosch, H., Langs, G.: Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. European Radiology Experimental **4**(1), 1–13 (2020)
12. Hu, J., Li, Y., Lin, L., Chen, Y.W.: Integrating spatial prior adapter for enhancing sam performance in medical image segmentation. In: 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE). pp. 20–23. IEEE (2023)
13. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3942–3951 (2021)
14. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al.: Segment anything model for medical images? Medical Image Analysis **92**, 103061 (2024)
15. 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)
16. 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)
17. Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., Hong, Q.: Lvit: language meets vision transformer in medical image segmentation. IEEE transactions on medical imaging (2023)
18. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
19. Ma, J., Wang, B.: Segment anything in medical images. arXiv e-prints pp. arXiv–2304 (2023)
20. Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)
21. 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)
22. 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. pp. 234–241. Springer (2015)
23. Tomar, N.K., Jha, D., Bagci, U., Ali, S.: Tganet: Text-guided attention for improved polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 151–160. Springer (2022)
24. Wu, J., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y., Arbel, T.: Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
25. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)
26. Zhong, Z., Tang, Z., He, T., Fang, H., Yuan, C.: Convolution meets lora: Parameter efficient finetuning for segment anything model. arXiv preprint arXiv:2401.17868 (2024)

27. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging **39**(6), 1856–1867 (2019)