Beyond Adapting SAM: Towards End-to-End Ultrasound Image Segmentation via Auto Prompting

Xian Lin[1], Yangyang Xiang[1], Li Yu[1], and Zengqiang Yan[1](✉)

School of Electronic Information and Communications, Huazhong University of Science and Technology
{xianlin, xyy_2001, hustlyu, z_yan}@hust.edu.cn

Abstract. End-to-end medical image segmentation is of great value for computer-aided diagnosis dominated by task-specific models, usually suffering from poor generalization. With recent breakthroughs brought by the segment anything model (SAM) for universal image segmentation, extensive efforts have been made to adapt SAM for medical imaging but still encounter two major issues: 1) severe performance degradation and limited generalization without proper adaptation, and 2) semi-automatic segmentation relying on accurate manual prompts for interaction. In this work, we propose SAMUS as a universal model tailored for ultrasound image segmentation and further enable it to work in an end-to-end manner denoted as AutoSAMUS. Specifically, in SAMUS, a parallel CNN branch is introduced to supplement local information through cross-branch attention, and a feature adapter and a position adapter are jointly used to adapt SAM from natural to ultrasound domains while reducing training complexity. AutoSAMUS is realized by introducing an auto prompt generator (APG) to replace the manual prompt encoder of SAMUS to automatically generate prompt embeddings. A comprehensive ultrasound dataset, comprising about 30k images and 69k masks and covering six object categories, is collected for verification. Extensive comparison experiments demonstrate the superiority of SAMUS and AutoSAMUS against the state-of-the-art task-specific and SAM-based foundation models. We believe the auto-prompted SAM-based model has the potential to become a new paradigm for end-to-end medical image segmentation and deserves more exploration. Code and data are available at https://github.com/xianlin7/SAMUS.

Keywords: SAM · Auto prompt · Foundation model · Medical image segmentation.

1 Introduction

Medical image segmentation, a crucial technology to discern and highlight specific organs, tissues, and lesions within medical images, serves as an integral component of computer-aided diagnosis systems [1]. Numerous deep-learning

models have been proposed for medical image segmentation, showcasing substantial potential [2,3]. However, these models are tailored for specific objects and necessitate training new model parameters when applied to other objects, resulting in great inconvenience for clinical applications with diverse tasks.

Segment anything model (SAM), serving as a versatile foundation model for vision segmentation, has garnered considerable acclaim owing to its remarkable segmentation capabilities across diverse objects and robust zero-shot generalization capacity [4]. According to user prompts, including points, bounding boxes, and coarse masks, SAM is capable of segmenting the corresponding objects. Therefore, through simple prompting, SAM can be effortlessly adapted to various segmentation applications. This paradigm enables the integration of multiple individual medical image segmentation tasks into a unified framework (*i.e.*, a universal model), greatly facilitating clinical deployment [5].

Despite constructing the largest dataset to date (*i.e.*, SA-1B), SAM encounters a rapid performance degradation in the medical domain due to the scarcity of reliable clinical annotations [5]. Some foundation models have been proposed to adapt SAM to medical image segmentation by tuning SAM on medical datasets [6,8]. However, the same as SAM, they perform a no-overlap $16\times$ tokenization on the input images before feature modeling, which destroys the local information crucial for identifying small targets and boundaries, making them struggle to segment clinical objects with complex/threadlike shapes, weak boundaries, small sizes, or low contrast. In addition, these SAM-based models require manually providing task-related prompts to generate the corresponding masks, leading to a semi-automatic segmentation pipeline. Such a paradigm is inflexible when dealing with certain clinical tasks.

In this paper, we present SAMUS first to transfer the strong feature representation ability of SAM to the domain of medical image segmentation, and then extend the trained SAMUS into an automatic version (*i.e.*, AutoSAMUS) to flexibly handle various downstream segmentation tasks. Specifically, SAMUS inherits the ViT image encoder, prompt encoder, and mask decoder of SAM, with tailored designs to the image encoder. First, we shorten the sequence length of the ViT branch by reducing the required input size to lower the computational complexity. Then, a feature adapter and a position adapter are developed to fine-tune the ViT image encoder from natural to medical domains. To complement local (*i.e.*, low-level) information in the ViT image encoder, we introduce a parallel CNN-branch image encoder, running alongside the ViT-branch and propose a cross-branch attention module to enable each patch in the ViT-branch to assimilate local information from the CNN-branch. A large ultrasound dataset called US30K is constructed to comprehensively train and evaluate the efficacy of SAMUS. After obtaining the trained SAMUS, it is expected to run in an end-to-end manner for downstream specific tasks. Therefore, we extend SAMUS into AutoSAMUS by introducing an auto prompt generator (APG) with learnable task tokens to replace the manual prompt encoder of SAMUS for generating task-related prompt embeddings. Experimental results demonstrate that SAMUS outperforms the state-of-the-art (SOTA) task-specific and universal seg-

mentation approaches. More importantly, based on the trained SAMUS, adjusting AutoSAMUS on specific tasks can realize end-to-end automatic segmentation and achieve better segmentation performance compared to SOTA task-specific methods. This indicates that developing auto-prompted SAM-based models is promising as a new end-to-end segmentation paradigm.

2 Related Works

Adapt SAM to Medical Image Segmentation. SAM has demonstrated remarkable performance in natural images but struggles with some medical image segmentation tasks, especially on objects with complex shapes, blurred boundaries, small sizes, or low contrast [5]. To bridge this gap and enable SAM to adapt effectively to the medical image domain, several methods have been proposed by applying vision tuning techniques to SAM. Specifically, MedSAM trains SAM on medical images by freezing the prompt encoder, focusing on tuning the image encoder and mask decoder [6]. SAMed applies the low-rank-based (LoRA) strategy on the image encoder to tune SAM at a lower computational cost, making it more feasible for medical image segmentation [7]. MSA adopts down-ReLU-up adapters on the ViT image encoder and mask decoder to introduce medical information [8]. Compared to current SAM-based universal models, the proposed SAMUS focuses more on complementing local features and realizing end-to-end automatic segmentation.

Prompts in SAMs. Vanilla SAM generates task-related masks under the driven of precise spatial prompts, *e.g.*, points, bounding boxes, and masks. To automatically obtain these spatial prompts, some methods introduce separate input-related networks. Specifically, AdapterShadow, SAC, and UV-SAM use EfficientNet, U-Net, and SegFormer respectively to generate coarse masks for making spatial prompts [9,10,11]. The introduced parameters of such separate networks are on the same level as task-specific methods, making the universal model cumbersome. Polyp-SAM++ uses Grounding DINO to generate bounding box prompts from the text prompts [12]. Adaptive SAM and SP-SAM encode text prompts into prompt embeddings by CLIP [13,14]. Although these approaches can effectively utilize text information, there is a lack of text-image data in medical scenes to tune the text encoders trained on natural scenes. SurgicalSAM proposes a prototype-based class prompt encoder to generate the dense and sparse prompt embeddings [15]. Auto-prompting SAM develops an auto-prompt encoder by constructing Up-Down full convolution layers [16]. These prompt encoders are deeply coupled with the mask decoder, causing difficulty in constructing robust feature representations through multi-objective learning for SAM-based models. Comparatively, the proposed APG is a lightweight and independent module and highly extendable to other SAM-based foundation models.

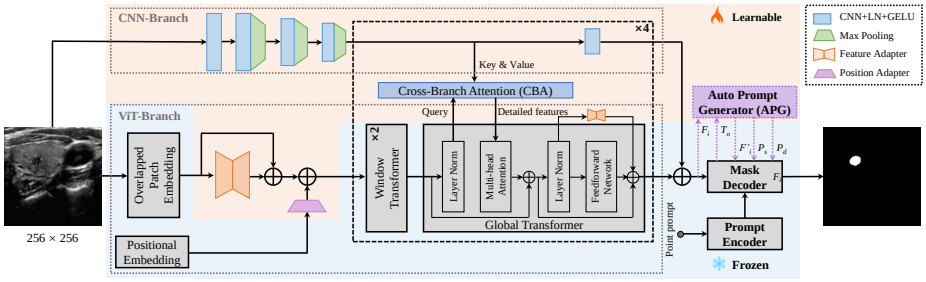

Fig. 1: Overview of the proposed SAMUS. APG represents the expansion module for extending SAMUS into AutoSAMUS.

3 Method

The Overall Architecture of SAMUS. As depicted in Fig. 1, the overall architecture of SAMUS is inherited from SAM, retaining the structure and parameters of the prompt encoder and the mask decoder without any adjustment. Comparatively, the image encoder is carefully modified to address the challenges of inadequate local features and excessive computational memory consumption, making it more clinically friendly. Major modifications include reducing the required input size, overlapping the patch embedding, introducing adapters to the ViT branch, adding a CNN branch, and introducing cross-branch attention (CBA). Specifically, the input spatial resolution is scaled down from 1024×1024 pixels to 256×256 pixels, resulting in a substantial reduction in GPU memory cost due to the shorter input sequence in transformers. The overlapped patch embedding uses the same parameters as the patch embedding in SAM while its patch stride is half to the original stride, well keeping the information from patch boundaries. Adapters in the ViT branch include a position adapter and five feature adapters. The position adapter is to accommodate the global position embedding in shorter sequences due to the smaller input size. The first feature adapter follows the overlapped patch embedding to align input features with the required feature distribution of the pre-trained ViT image encoder. The remaining feature adapters are attached to the residual connections of the feed-forward network in the global transformer to fine-tune the pre-trained image encoder. In terms of the CNN branch, it is parallel to the ViT branch, providing complementary local information to the latter through the CBA module, which takes the ViT-branch features as the query and builds global dependency with features from the CNN branch. It should be noted that CBA is only integrated into each global transformer. Finally, the outputs of both the two branches are combined as the final encoded image embedding F_i of SAMUS.

Adapters in the ViT Branch. To facilitate the generalization of the trained image encoder (*i.e.*, the ViT branch) of SAM to smaller input sizes and the medical image domain, we introduce a position adapter and five feature adapters. These adapters can effectively tune the ViT branch while only requiring much

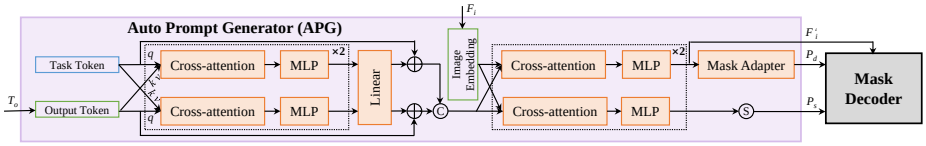

Fig. 2: Details of the auto prompt generator.

fewer parameters. Specifically, the position adapter is responsible for adjusting the positional embedding to match the resolution of the embedded sequence. It begins by downsampling the positional embedding through max pooling with the stride and kernel size as 2, achieving the same resolution as the embedded sequence. Then, a convolution operation with a kernel size of 3×3 is applied to tune the position embedding, further aiding the ViT branch in better handling smaller inputs. All feature adapters have the same structure that comprises three components: a down linear projection, an activation function, and an up linear projection. The procedure of each feature adapter can be formulated as:

$$\mathcal{A}(x) = \mathcal{G}(xE_d)E_u, \tag{1}$$

where $\mathcal{G}$ represents the GELU activation founction, $E_d \in \mathbb{R}^{d \times \frac{d}{4}}$ and $E_u \in \mathbb{R}^{\frac{d}{4} \times d}$ are the projection matrices, d is the dimension of the feature embedding. Through these simple operations, feature adapters enable the ViT branch to better adapt to the feature distribution of medical image domains.

The CNN Branch. The CNN branch consists of a series of convolution-pooling blocks connected in sequence. Specifically, the inputs pass through a single convolution block initially, followed by being processed through three convolution-pooling blocks. Then, the feature maps in the CNN branch share the same spatial resolution as those of the ViT branch. In the rest part of the CNN branch, such single convolution blocks are repeated four times in sequence. Each single convolution block contains a convolution with a kernel size of 3×3 and each convolution-pooling block contains a max pooling with the stride and kernel size as 2 and a single convolution block. More details are illustrated in Fig. 1. This minimalist and lightweight design of the CNN branch is to prevent overfitting during training.

Cross-branch Attention. The cross-branch attention (CBA) module creates a bridge between the CNN branch and the ViT branch to further complement missing local features with the ViT branch. For a pair of feature maps from the ViT branch F_v and the CNN branch F_c, cross-branch attention in the single head can be formulated as:

$$\mathcal{F}(F_v, F_c) = (\sigma(\frac{F_v E_q (F_c E_k)^T}{\sqrt{d_m}}) + R)(F_c E_v), \tag{2}$$

where σ represents the Softmax function. $E_q \in \mathbb{R}^{d \times d_m}$, $E_k \in \mathbb{R}^{d \times d_m}$, and $E_v \in \mathbb{R}^{d \times d_m}$ are the learnable weight matrices used to project F_c and F_v to different feature subspaces. $R \in \mathbb{R}^{hw \times hw}$ is the relative position embedding, and d_m is

Table 1: Universality comparison of SAMUS and SOTA foundation models on segmenting thyroid nodule (TN3K), breast cancer (BUSI), left ventricle (CAMUS-LV), myocardium (CAMUS-MYO), and left atrium (CAMUS-LA).

Method	TN3K		BUSI		CAMUS-LV		CAMUS-MYO		CAMUS-LA	
	Dice	HD	Dice	HD	Dice	HD	Dice	HD	Dice	HD
SAM [4]	29.59	134.87	54.01	82.39	28.18	196.64	29.42	184.10	17.28	193.70
MedSAM [6]	71.09	42.91	77.75	34.26	87.52	15.28	76.07	25.72	88.06	15.70
SAMed [7]	80.40	31.29	74.82	34.60	87.67	13.24	82.60	19.48	90.92	12.60
MSA [8]	82.67	29.15	81.66	28.87	90.95	**11.29**	82.47	19.28	91.80	**11.59**
SAMUS	**83.05**	**28.82**	**84.54**	**27.24**	**91.13**	11.76	**83.11**	**18.99**	**92.00**	12.08

the dimension of CBA. The final output of CBA is the linear combination of g such single-head attention.

AutoSAMUS. AutoSAMUS, an end-to-end automatic segmentation framework extended from SAMUS, is realized by replacing the manual prompt encoder of SAMUS with APG. As depicted in Fig. 2, the inputs of APG consist of the output tokens $T_o \in \mathbb{R}^{5\times d}$ and the image embedding F_i, where T_o is the frozen parameters extracted from the mask decoder and F_i is the output of the image encoder. To indicate the segmentation task, APG introduces learnable task tokens $T_t \in \mathbb{R}^{k\times d}$ for automatically generating task-related prompt embeddings, where k is the number of task tokens. Firstly, the combination of cross-attention and multi-layer perceptron (MLP) is used to couple the task tokens and the output tokens, formulated as:

$$\mathcal{C}(T_t, T_o) = \mathbf{MLP}(\sigma(\frac{T_t W_q (T_o W_k)^T}{\sqrt{d}})(T_o W_v)), \tag{3}$$

where **MLP** consists of two linear layers. W_q, W_k, and $W_v \in \mathbb{R}^{d\times d}$ are the learnable weight matrices. Then, the updated task tokens are represented as:

$$T_{t_1} = \mathcal{C}(\mathcal{C}(T_t, T_o), \mathcal{C}(T_o, T_t))W + T_t, \tag{4}$$

where $W \in \mathbb{R}^{d\times d}$ is the projection matrix. Similarly, by swapping the positions of T_t and T_o in Eq. 4, the updated output tokens T_{o_1} can be calculated. Next, to adapt the image embedding to the task domain and make the task tokens aware of image information, we perform the combination operation $\mathcal{C}$ defined in Eq. 3 between the image embedding and the combined tokens $T = [T_{t_1}, T_{o_1}]$. After that, the updated image embedding and the combined tokens are represented as $F_i' = \mathcal{C}(\mathcal{C}(F_i, T), \mathcal{C}(T, F_i))$ and $T' = \mathcal{C}(\mathcal{C}(T, F_i), \mathcal{C}(F_i, T))$. Finally, based on T' and F_i', APG generates the sparse prompt embedding $P_s = T'[:k, :]$ and the dense prompt embedding $P_d = \mathcal{M}(F_i')$ to prompt the frozen mask decoder, where $\mathcal{M}$ represents the sequence of operation consisting of four single convolution blocks with channels of $\frac{d}{4}$, $\frac{d}{4}$, $\frac{d}{4}$, and d. In addition, the updated image embedding F_i' will replace the original image embedding to participate in mask decoding.

Table 2: Generalization comparison of foundation models on segmenting thyroid nodule (DDTI), breast cancer (UDIAT), and myocardium (HMC-QU)

Dataset	SAM [4]		MedSAM [6]		SAMed [7]		MSA [8]		SAMUS	
	Dice	HD	Dice	HD	Dice	HD	Dice	HD	Dice	HD
DDTI	25.57	116.23	57.94	51.77	61.64	**43.45**	62.24	46.49	**66.78**	44.35
UDIAT	49.18	104.43	61.49	50.80	72.56	30.82	76.24	**26.64**	**78.06**	26.91
HMC-QU	25.91	93.20	34.58	36.30	37.82	37.72	40.56	38.18	**56.77**	**25.21**

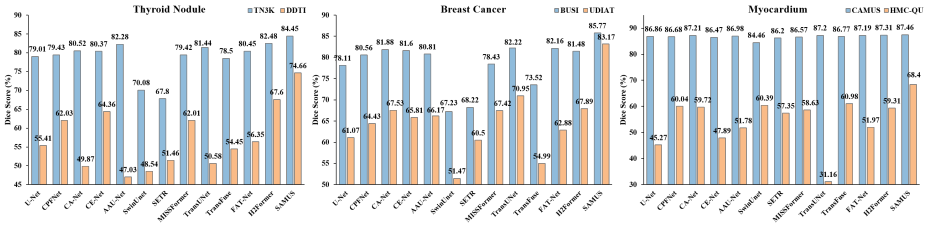

Fig. 3: Comparison between SAMUS and task-specific methods evaluated on seeable (marked in blue) and unseen datasets (marked in orange).

4 Experiments

Datasets. To comprehensively evaluate the effectiveness of SAMUS, we have constructed a large ultrasonic dataset named US30K as summarized in the supplementary material, containing data from seven publicly-available datasets, including TN3K [17], DDTI [18], TG3K [19], BUSI [20], UDIAT [21], CAMUS [22], and HMC-QU [23]. The data in TN3K and TG3K is partitioned into train, validation, and test sets following TRFE [17]. BUSI is randomly split into 7:1:2 for training, validation, and testing, respectively. CAMUS is divided into a train set and a test set first according to the challenge [22]. Then, we randomly select 10% patients from the train set to validate the model and the rest data as the final training data. To evaluate the generalization of different models, the other datasets in US30K are unseen during training. The comparison between SAMUS and other foundation models is conducted by training on the entire training set of US30K and evaluated on separate tasks. For a fair comparison, all foundation models are re-implemented and trained for 400 epochs under the same settings using the same single-point prompt. The comparison between SAMUS and SOTA task-specific approaches is conducted by training on each single dataset.

Compare SAMUS with SOTA Foundation Models. Comparison of universality and generalization among foundation models on US30K are summarized in Tables 1 and 2. Among comparison foundation models, MAS [8] is the best-performing model both in universality and generalization. Compared to MSA, SAMUS consistently achieves remarkable improvements across all subdatasets of US30K. It indicates the value of the CNN branch and the CBA module in

Table 3: Quantitative dice (%) comparison of AutoSAMUS and SOTA task-specific methods. AutoSAMUS- is the model when only APG is updated.

Method	U-Net	CPFNet	CANet	AAU-Net	MISSFormer	TransUNet	H2Former	AutoSAMUS-	AutoSAMUS
DDTI	76.71	78.38	75.79	78.16	78.11	77.52	77.92	78.89	**82.64**
UDIAT	82.64	82.37	82.68	81.46	82.89	81.15	82.26	84.17	**85.39**
HMC-QU	93.26	93.58	93.65	93.66	93.50	93.45	93.44	92.87	**94.10**

Table 4: Ablation study on different component combinations of SAMUS. F-Adapter and P-Adapter represent feature and position adapters respectively.

Components				TN3K		DDTI		BUSI		UDIAT	
CNN Branch	CBA	F-Adapter	P-Adapter	Dice	HD	Dice	HD	Dice	HD	Dice	HD
✗	✗	✗	✗	29.59	134.87	25.57	116.23	54.01	82.39	49.18	104.43
✓	✗	✗	✗	82.17	31.41	68.31	48.66	81.42	29.50	82.24	22.53
✓	✓	✗	✗	83.65	28.47	72.71	35.76	83.53	30.26	80.87	25.60
✗	✗	✓	✗	83.64	29.83	70.38	45.29	84.53	26.30	81.25	23.18
✗	✗	✗	✓	80.19	32.12	63.67	53.86	80.78	29.00	79.72	24.71
✓	✓	✓	✓	**84.45**	**28.22**	**74.66**	**21.03**	**85.77**	**25.49**	**83.17**	**21.25**

SAMUS for complementing local information which is crucial for medical image segmentation.

Compare SAMUS with SOTA Task-Specific Methods. As depicted in Fig. 3, 12 SOTA methods are included for comparison [2,3,24,25,26,27,28,29,30,31,32,33]. SAMUS surpasses the best comparison methods across all datasets including TN3K, BUSI, CAMUS-MYO, DDTI, UDIAT and HMC-QU with an average increase of 1.97%, 3.55%, 0.15%, 7.06%, 12.22% and 7.42% in Dice, respectively. It proves the necessity of adapting SAM to the medical image domain by SAMUS.

Compare AutoSAMUS with SOTA Task-Specific Methods. To compare AutoSAMUS with task-specific methods, we first load the trained parameters of SAMUS into AutoSAMUS. Then, we fine-tune APG and APG together with the learnable parts of SAMUS on three downstream tasks, the results are represented by AutoSAMUS- and AutoSAMUS respectively in Table 3. AutoSAMUS- surpasses the best comparison method on DDTI and UDIAT and approaches the best comparison method on HMC-QU, while AutoSAMUS outperforms the best comparison method on DDTI, UDIAT, and HMC-QU with an average increase of 4.26%, 2.5%, and 0.44% in Dice, respectively.

Effectiveness of each component in SAMUS. As summarized in Table 4, coupling any component of SAMUS can effectively improve the segmentation performance and generalization ability of SAM. Combining all four components, SAMUS achieves the best segmentation and generalization performance. 8

5 Conclusion

In this paper, we propose SAMUS, a universal foundation model derived from SAM, and its end-to-end version AutoSAMUS, for clinically-friendly and generalizable ultrasound image segmentation. Specifically, we present a CNN branch

image encoder, a feature adapter, a position adapter, and a cross-branch attention module to enrich the feature representations of ultrasound objects. Furthermore, to facilitate the clinical application of downstream tasks, we combine SAMUS with an auto prompt generator for automatic segmentation, which realizes a new end-to-end segmentation paradigm instead of relying on manual prompts as vanilla SAM. A large ultrasound image dataset US30K consisting of 30k+ images and 68k+ masks is constructed for evaluation and potential clinical usage. Extensive experiments demonstrate the outstanding performance of SAMUS and AutoSAMUS, outperforming SOTA both SAM-based medical foundation models and task-specific models.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62271220 and Grant 62202179, in part by the Natural Science Foundation of Hubei Province of China under Grant 2022CFB585, and in part by the Fundamental Research Funds for the Central Universities, HUST: 2024JYCXJJ032. The computation is supported by the HPC Platform of HUST.

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

References

1. Liu, X., Song, L., Liu, S., Zhang, Y.: A review of deep-learning-based medical image segmentation methods. Sustainability. **13**(3), 1224 (2021)
2. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W. M., Frangi, A.F. (eds.) MICCAI 2015, LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
3. Huang, X., Deng, Z., Li, D., Yuan, X., Fu, Y.: MISSFormer: An effective transformer for 2d medical image segmentation. IEEE Trans. Med. Imag. **42**(5), 1484–1494 (2022)
4. Kirillov, A., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
5. Huang, Y., et al.: Segment anything model for medical images? Med. Image Anal. **92**, 103061 (2024)
6. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nat. Commun. **15**(1), 654 (2024)
7. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)
8. Wu, J., et al.: Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
9. Jie, L., Zhang, H.: AdapterShadow: Adapting segment anything model for shadow detection. arXiv preprint arXiv:2311.08891 (2023)
10. Na, S., Guo, Y., Jiang, F., Ma, H., Huang, J.: Segment any cell: A SAM-based auto-prompting fine-tuning framework for nuclei segmentation. arXiv preprint arXiv:2401.13220 (2024)
11. Zhang, X., Liu, Y., Lin, Y., Liao, Q., Li, Y.: UV-SAM: Adapting segment anything model for urban village identification. arXiv preprint arXiv:2401.08083 (2024)

12. Biswas, R.: Polyp-sam++: Can a text guided sam perform better for polyp segmentation?. arXiv preprint arXiv:2308.06623 (2023)
13. Paranjape, J. N., Nair, N. G., Sikder, S., Vedula, S. S., Patel, V. M.: Adaptivesam: Towards efficient tuning of sam for surgical scene segmentation. arXiv preprint arXiv:2308.03726 (2023)
14. Yue, W., et al.: Part to whole: Collaborative prompting for surgical instrument segmentation. arXiv preprint arXiv:2312.14481 (2023)
15. Yue, W., Zhang, J., Hu, K., Xia, Y., Luo, J., Wang, Z.: Surgicalsam: Efficient class promptable surgical instrument segmentation. arXiv preprint arXiv:2308.08746 (2023)
16. Li, C., Khanduri, P., Qiang, Y., Sultan, R. I., Chetty, I., Zhu, D.: Auto-prompting sam for mobile friendly 3d medical image segmentation. arXiv preprint arXiv:2308.14936 (2023)
17. Gong, H., Chen, J., Chen, G., Li, H., Li, G., Chen, F.: Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. Comput. Biol. Med. **155**, 106389 (2023)
18. Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., Romero, E.: An open access thyroid ultrasound image database. In: 10th International Symposium on Medical Information Processing and Analysis, pp. 188–193 (2015)
19. Wunderling, T., Golla, B., Poudel, P., Arens, C., Friebe, M., Hansen, C.: Comparison of thyroid segmentation techniques for 3D ultrasound. In: Image Processing 2017, pp. 346–352 (2017)
20. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in Brief. **28**, 104863 (2020)
21. Yap, M. H., et al.: Breast ultrasound region of interest detection and lesion localisation. Artif. Intell. in Med. **107**, 101880 (2020)
22. Leclerc, S., et al.: Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE Trans. Med. Imag. **38**(9), 2198–2210 (2019)
23. Kiranyaz, S., et al.: Left ventricular wall motion estimation by active polynomials for acute myocardial infarction detection. IEEE Access. **8**, 210301–210317 (2020)
24. Feng, S., et al.: CPFNet: Context pyramid fusion network for medical image segmentation. IEEE Trans. Med. Imag. **39**(10), 3008-3018 (2020)
25. Gu, R., et al.: CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Trans. Med. Imag. **40**(2), 699–711 (2020)
26. Gu, Z., et al.: Ce-net: Context encoder network for 2d medical image segmentation. IEEE Trans. Med. Imag. **38**(10), 2281–2292 (2019)
27. Chen, G., Li, L., Dai, Y., Zhang, J., Yap, M. H.: AAU-net: an adaptive attention U-net for breast lesions segmentation in ultrasound images. IEEE Trans. Med. Imag. **42**(5), 1289–1300 (2023)
28. Cao, H., et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision, pp. 205–218 (2022)
29. Zheng, S., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6881–6890 (2021)
30. Chen, J., et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
31. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021, LNCS, vol. 12901, pp. 14–24. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_2

32. Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., Wen, Z.: FAT-Net: Feature adaptive transformers for automated skin lesion segmentation: Medical Image Anal. **76**, 102327 (2022)
33. He, A., Wang, K., Li, T., Du, C., Xia, S., Fu, H.: H2former: An efficient hierarchical hybrid transformer for medical image segmentation. IEEE Trans. Med. Imag. **42**(9), 2763–2775 (2023)