Supplementary Materials for I-MedSAM: Implicit Medical Image Segmentation with Segment Anything

Xiaobao Wei[1,2,3,*], Jiajun Cao[1,4,*], Yizhu Jin[1,5], Ming Lu[1], Guangyu Wang[6], and Shanghang Zhang[1,†]

[1]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University [2]University of Chinese Academy of Sciences [3]Institute of Software, Chinese Academy of Sciences [4]Xi'an Jiaotong University [5]Beihang University [6]Beijing University of Posts and Telecommunications
weixiaobao0210@gmail.com

A Overview

The supplementary material encompasses the subsequent components.

- Additional related work
- Supplementary experiments
  - Cross-resolution experiment on different target resolutions
  - Cross-domain experiment on multi-class organ segmentation
  - Cross-domain experiment on large-scale medical datasets
  - Ablation study on different dropout probability
  - Ablation study on different LoRA ranks
- Computational complexity of inference
- Additional visualization results
  - Visualization for cross-resolution
  - Visualization for cross-domain
- Limitation

B Additional related work

Spectral Representation. In the context of medical image analysis, the emphasis on textual information over edge information aligns with findings that deep neural networks tend to bias towards learning low-frequency representations [14, 16]. To better leverage high-frequency information, numerous studies have extensively investigated the integration of spectral representation into deep neural networks, such as FFT, DCT, and Wavelet. Among these spectral representations, the FFT-based frequency representation emerges as particularly prevalent [13, 15, 21, 25]. The use of FFT-based frequency representation not

* Equal Contribution.
† Corresponding Author.

only establishes a robust foundation for modeling but also facilitates various operations, harnessing the advantages of both spectral and spatial representations [18,19]. While recent efforts have focused on developing adapters for SAM, there is a notable oversight in integrating information from the frequency domain to achieve accurate segmentation boundaries. To the best of our knowledge, we are the first to introduce an adapter for SAM that embeds frequency features, enhancing high-frequency modeling for more precise boundary delineation.

C Supplementary experiments

C.1 Cross-resolution experiment on different target resolutions

As shown in Fig. 1 and Fig. 2, for a fixed source input resolution of 384×384, with the target resolution ranges from 16×16 to 2048×2048, I-MedSAM consistently achieves high-precision segmentation results across a broad spectrum of target resolutions, particularly within the range of 64×64 to 2048×2048. It notably demonstrates superior performance in scaling to higher resolutions, all while preserving minimal loss in segmentation accuracy.

C.2 Cross-domain experiment on multi-class organ segmentation

To further showcase the generalization ability of our method across organs of varying sizes, we implement cross-domain segmentation experiments from BCV [11] to AMOS [8] (a substantially larger dataset compared to BCV) for multi-class organ segmentation in Tab. 1. We compare against state-of-the-art discrete methods like nnUNet [6], MedSAM [12], and the implicit method represented by IOSNet [9]. In this experiment, we exclude SwIPE [23] as one of our major comparison baselines, which is also an implicit method, due to its limited reporting of segmentation results only for the liver class in the same setting, and its code is not available for reproduction. The experimental results indicate that models leveraging SAM [10] as the backbone exhibit superior generalization capabilities across different organs. Compared to IOSNet, an implicit method utilizing ResNet-based backbone, SAM-based methods with ViT as the backbone demonstrate superior performance in terms of generalization.

Table 1: Cross-domain experiment results of multi-class organ segmentation from BCV to AMOS (Dice %).

Classes	Spleen	Right Kidney	Left Kidney	Gall Bladder	Esophagus	Stomach	Arota
IOSNet [9]	72.19	86.66	79.21	31.69	39.26	59.17	71.16
nnUNet [6]	73.82	46.14	57.83	47.06	38.94	62.76	82.80
MedSAM [12]	81.15	85.22	84.22	77.52	**82.29**	78.92	86.91
I-MedSAM(Ours)	**90.44**	**91.10**	**90.78**	**82.88**	71.78	**85.27**	**88.80**

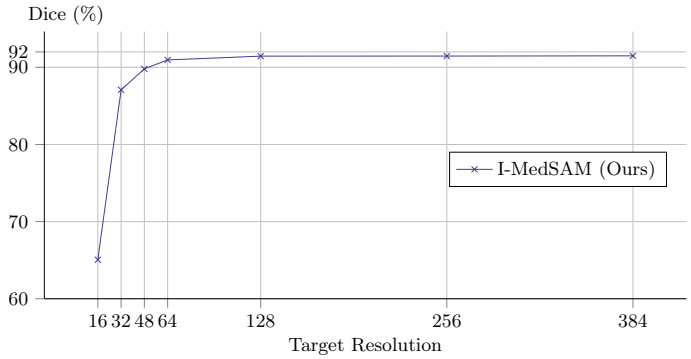

Fig. 1: Experiments on **lower** target resolutions on Kvasir-Sessile [7] dataset.

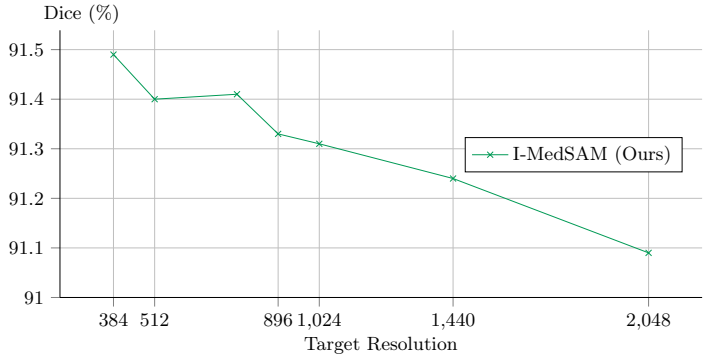

Fig. 2: Experiments on **higher** target resolutions on Kvasir-Sessile [7] dataset.

C.3 Cross-domain experiment on large-scale medical datasets

To further evaluate the cross-domain performance on larger scale datasets, we select CTooth+ [4], a large-scale dental cone beam computed tomography dataset. CTooth+ comprises 22 fully labeled and 146 unlabeled volumes. We divide the 3D volumes into slices, retaining those slices where the sum of pixel-wise 0-1 labels exceeded 100 (following a data processing method similar to that proposed in [5]). This process results in 1600 training samples and 400 test samples. Additionally, we utilize the Panoramic radiography database [17], consisting of 598 panoramic radiographs, to assess the cross-domain capability of the model. We partition the training and testing sets at an 80:20 ratio, and the cross-domain experiment is conducted using pre-trained weights from CTooth+ on the testing set. Experiments on both datasets take 256×256 images as inputs.

Since SwIPE's code is not publicly available, we select IOSNet [9] as a representative of the implicit approach. As depicted in Tab. 2, our model significantly outperforms MedSAM [12] and IOSNet on Ctooth+, although it slightly underperforms compared to nnUNet [6]. However, our model demonstrates superior

Table 2: Evaluate cross-domain performance on a larger scale dataset with Dice(%).

Settings	CTooth+	CTooth+ → Panoramic
IOSNet [9]	79.68	65.47
MORSE [20]	89.21	74.17
SAMed [22]	89.73	84.93
MedSAM [12]	84.31	74.14
H-SAM [3]	91.29	81.08
SAM-Med2D [2]	87.69	82.06
nnUNet [6]	**92.49**	85.47
I-MedSAM(Ours)	91.34	**87.01**

cross-domain capabilities, indicating its proficiency in handling various dataset sizes. This highlights the practical relevance of our model in clinical medical image segmentation across different domains.

C.4 Ablation study on different dropout probability

Table 3: Effect on different dropout probability.

Setting (x100%)	0.0	0.1	0.4	0.5	0.6	0.9
Dice (%)	87.74	90.12	90.89	**91.34**	90.59	89.75

To further investigate the impact of dropout, we conduct experiments with different dropout probability settings. The results are presented in Tab. 3. It is crucial to select an appropriate dropout probability for uncertainty-guided sampling to ensure robust training. A smaller dropout probability may overly smooth the variance distribution, making it challenging for the coarse INR to identify difficult samples for the fine INR to refine. Conversely, a larger dropout probability may introduce instability to the training process. For a new dataset, we recommend initially setting this hyperparameter to 0.5 and adjusting it based on specific circumstances as needed.

C.5 Ablation study on different LoRA ranks

Table 4: Ablation study on LoRA ranks.

LoRA Ranks	8	6	4
Dice (%)	90.29	90.41	**91.49**
HD	12.41	12.39	**11.59**

Tab. 4 illustrates the impact of different values of LoRA ranks in I-MedSAM. When LoRA ranks is set to 4, it exhibits optimal performance while minimizing training parameters. This demonstrates the effectiveness of this parameter-efficient fine-tuning technique.

D Computational complexity of inference

We outline the inference stage implementation and the computational cost. We start with combined features of $\mathbb{R}^{HW\times C}$ from SAM's image encoder and prompt encoder, concatenating the original positional embedding with grid coordinates as $z^p \in \mathbb{R}^{HW\times(C+C^p)}$. The coarse INR Dec_c processes this into $\mathbb{R}^{HW\times(C'+C^o)}$. Using UGS, we sample the Top-k percent of feature points with the highest variance T times based on the coarse prediction result $\mathbb{R}^{HW\times C^o}$, which has a computational cost of $\mathcal{O}(HW\log(HW))$, resulting in $z^s \in \mathbb{R}^{kHW\times C'}$. The fine INR Dec_f then refines the selected features into $\mathbb{R}^{HW\times C^o}$. Both Dec_c and Dec_f are MLPs, with computational cost related to the intermediate feature dimensions, represented as $\mathcal{O}(HW\sum_{i=1}^{N-1} D_i \cdot D_{i+1})$, where N is the number of linear layers and D_i is the dimension of each layer. Detailed computational evaluation for MLPs is case-specific and omitted here for simplicity.

E Additional visualization results

E.1 Visualization for cross-resolution

As illustrated in Fig. 3 and Fig. 4, we conduct cross-resolution experiments on Kvasir-Sessile, with segmentation boundaries highlighted in green lines. The figures demonstrate that I-MedSAM consistently maintains accurate boundaries across different resolutions. In contrast, baselines employing discrete representations like nnUNet [6] struggle to segment target objects accurately when presented with inputs of varying resolutions. Additionally, directly interpolating ground truth segmentation maps often results in either blurred boundaries or sparse segmentation maps.

E.2 Visualization for cross-domain

As depicted in Fig. 5 and Fig. 6, we conduct cross-domain experiments from Kvasir-Sessile to CVC [1] and from BCV [11] to AMOS [8] datasets. Our approach, I-MedSAM, along with baseline methods, is trained on the source domain and directly tested on the target domain. Following SwIPE [23], we specifically compare liver segmentation results for the BCV to AMOS transition. From the figures, it is evident that I-MedSAM achieves superior segmentation maps and demonstrates the best generalization capability.

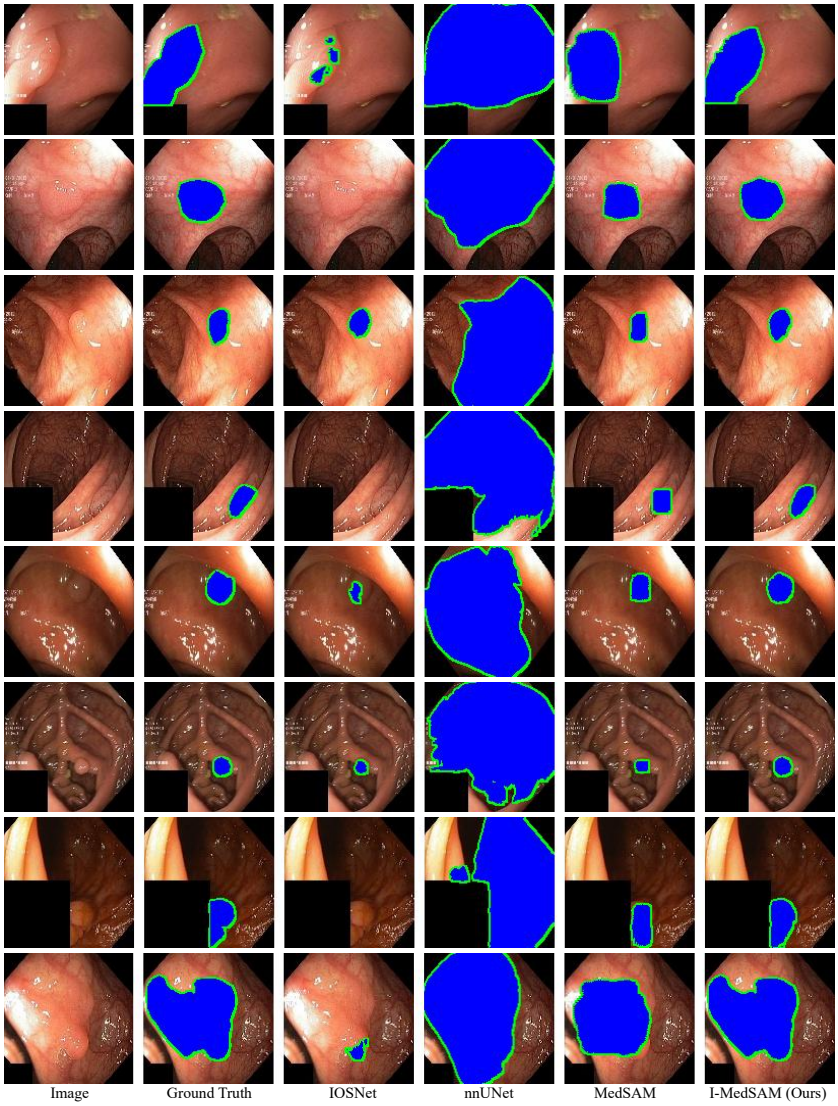

Fig. 3: Qualitative comparison for cross-resolution experiment from 384×384 to 128×128. The blurring in the image is normal because we zoom in on a low-resolution image directly in latex. Best viewed in colors.

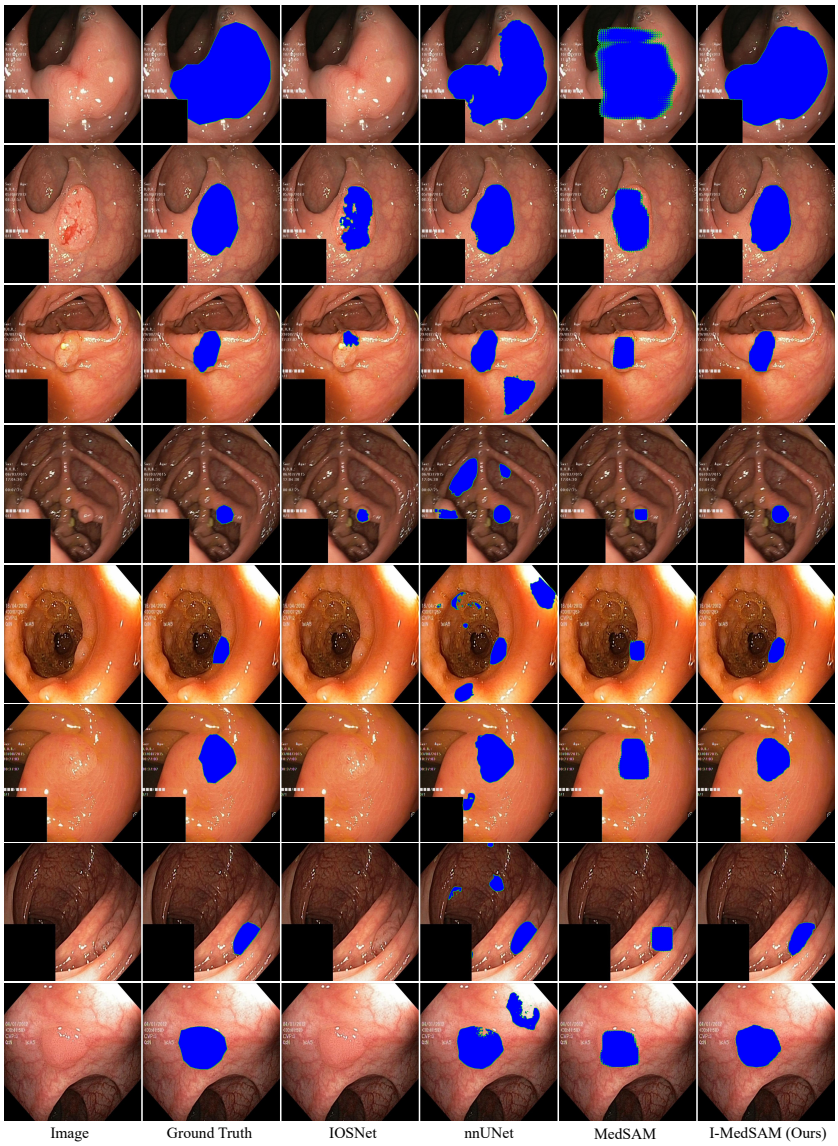

Fig. 4: Qualitative comparison for cross-resolution experiment from 384×384 to 896×896. Please zoom in for more boundary details.

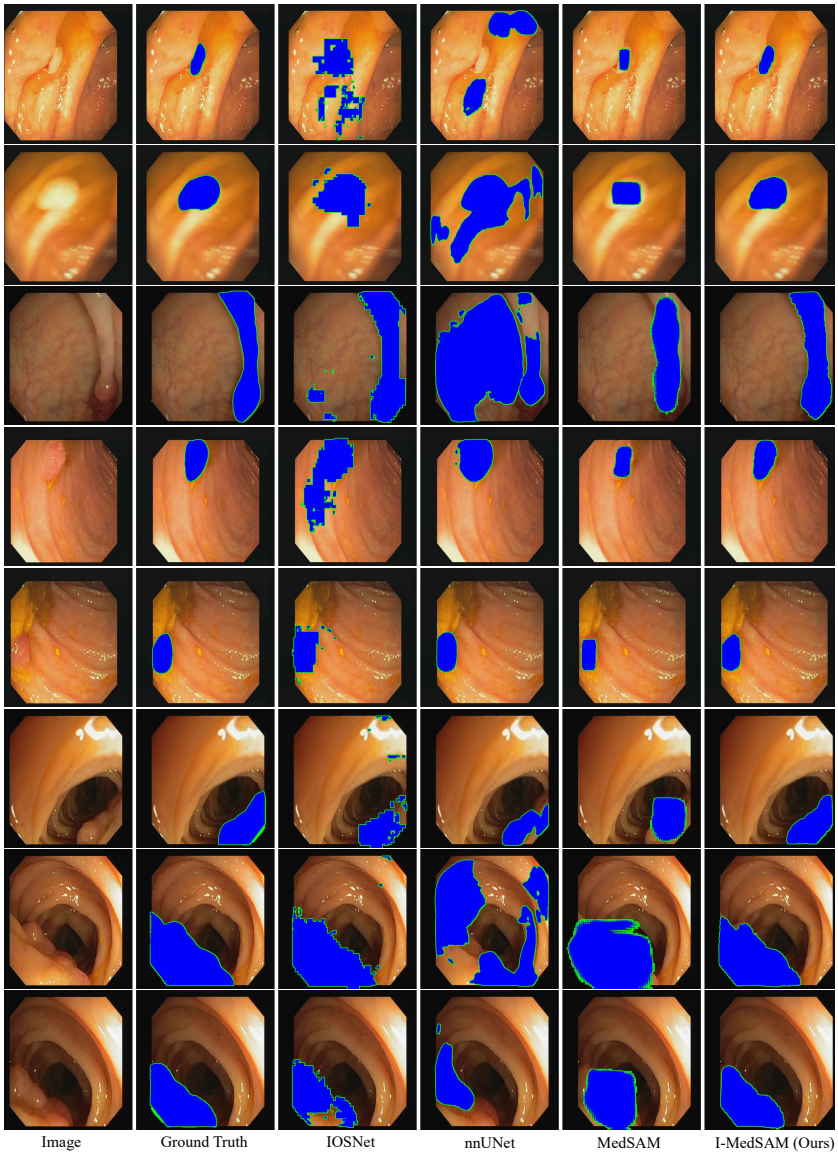

Fig. 5: Qualitative comparison for cross-domain from Kvasir-Sessile to CVC dataset. Best viewed in colors. Please zoom in for more boundary details.

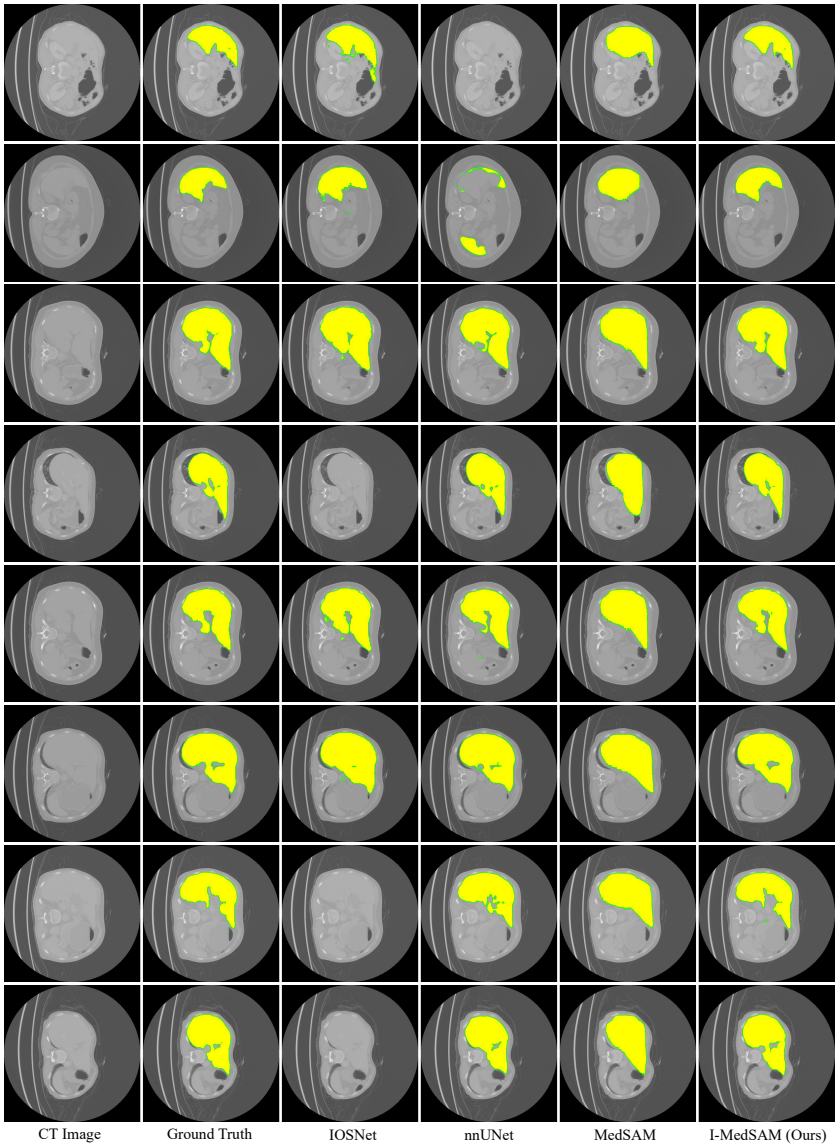

Fig. 6: Qualitative comparison for cross-domain from BCV to AMOS dataset on liver segmentation. Best viewed in colors. Please zoom in for more boundary details.

F Limitation

While SAM has been equipped with various adapters to address diverse tasks, the simultaneous handling of medical images from multiple domains, including MRI and CT, remains a formidable challenge in the field of medical image processing [24]. A potential strategy for improvement is the extension of I-MedSAM with a more universal adapter capable of generalizing across different modalities. Despite these challenges, I-MedSAM has effectively showcased the utility of leveraging Implicit Neural Representations (INRs) to enhance SAM's adaptability, demonstrating promising outcomes in aligning with out-of-distribution data for specific tasks.

References

1. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics **43**, 99–111 (2015)
2. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
3. Cheng, Z., Wei, Q., Zhu, H., Wang, Y., Qu, L., Shao, W., Zhou, Y.: Unleashing the potential of sam for medical adaptation via hierarchical decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3511–3522 (2024)
4. Cui, W., Wang, Y., Li, Y., Song, D., Zuo, X., Wang, J., Zhang, Y., Zhou, H., Chong, B.s., Zeng, L., et al.: Ctooth+: A large-scale dental cone beam computed tomography dataset and benchmark for tooth volume segmentation. In: MICCAI Workshop on Data Augmentation, Labelling, and Imperfections. pp. 64–73. Springer (2022)
5. 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)
6. Isensee, F., Jaeger, P., Kohl, S., 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)
7. Jha, D., Smedsrud, P.H., Johansen, D., de Lange, T., Johansen, H.D., Halvorsen, P., Riegler, M.A.: A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation. IEEE Journal of Biomedical and Health Informatics **25**(6), 2029–2040 (2021)
8. 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:2206.08023 (2022)
9. Khan, M., Fang, Y.: Implicit neural representations for medical imaging segmentation. In: MICCAI (2022)
10. 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)
11. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI

Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. vol. 5, p. 12 (2015)
12. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 1–9 (2024)
13. Mao, X., Liu, Y., Liu, F., Li, Q., Shen, W., Wang, Y.: Intriguing findings of frequency selection for image deblurring. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1905–1913 (2023)
14. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
15. Prabhu, A., Farhadi, A., Rastegari, M., et al.: Butterfly transform: An efficient fft based neural architecture design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12024–12033 (2020)
16. Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., Courville, A.: On the spectral bias of neural networks. In: International Conference on Machine Learning. pp. 5301–5310. PMLR (2019)
17. Román, J.C.M., Fretes, V.R., Adorno, C.G., Silva, R.G., Noguera, J.L.V., Legal-Ayala, H., Mello-Román, J.D., Torres, R.D.E., Facon, J.: Panoramic dental radiography image enhancement using multiscale mathematical morphology. Sensors **21**(9), 3110 (2021)
18. Tang, X., Peng, J., Zhong, B., Li, J., Yan, Z.: Introducing frequency representation into convolution neural networks for medical image segmentation via twin-kernel fourier convolution. Computer Methods and Programs in Biomedicine **205**, 106110 (2021)
19. Xu, Z.Q.J., Zhang, Y., Xiao, Y.: Training behavior of deep neural network in frequency domain. In: Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part I 26. pp. 264–274. Springer (2019)
20. You, C., Dai, W., Min, Y., Staib, L., Duncan, J.S.: Implicit anatomical rendering for medical image segmentation with stochastic experts. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 561–571. Springer (2023)
21. Zhang, D., Ouyang, J., Liu, G., Wang, X., Kong, X., Jin, Z.: Ff-former: Swin fourier transformer for nighttime flare removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2823–2831 (2023)
22. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)
23. Zhang, Y., Gu, P., Sapkota, N., Chen, D.Z.: Swipe: Efficient and robust medical image segmentation with implicit patch embeddings. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 315–326. Springer (2023)
24. Zhang, Y., Jiao, R.: How segment anything model (sam) boost medical image segmentation? arXiv preprint arXiv:2305.03678 (2023)
25. Zhu, Q., Li, P., Li, Q.: Attention retractable frequency fusion transformer for image super resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1756–1763 (2023)