ASAM: ANATOMY-ENCODED SEGMENT ANYTHING MODEL FOR MEDICAL IMAGES

Jiajing Zhang, Haotian Guan, Wei-Ning Lee

Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong

2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI) | 979-8-3315-2052-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/ISBI60581.2025.10981134

ABSTRACT

SAM is a foundation model for image segmentation, aiming to segment any target in any scene. MedSAM is its pre-trained version on medical images and can be generalized to different imaging modalities. However, current studies have overlooked the biggest difference between natural and medical images: anatomical features pertaining to biological tissues. Real-world medical imaging can be conceptualized as a convolution of the system impulse response with its interrogated medium. We herein propose an Anatomy-encoded SAM (ASAM), featuring an unsupervised cross-attention to emulate the imaging process, where the values of full-image, anatomy, and modality characteristics constrain each other in every attention operation via convolution and deconvolution in Fourier domain. Specifically, anatomy values are extracted from full-image values by deconvolution with modality values. ASAM focuses on CT, MR, US, X-ray, and PET modalities, and is trained on 37 large-scale public datasets. ASAM achieves a high average Dice of 91.46%, thereby being an advanced foundation model for medical image segmentation.

***Index Terms*—** Segmentation, Foundation model, Anatomy, Deep learning

1. INTRODUCTION

Medical image segmentation plays an important role in diagnosis, treatment planning, and image-guided surgery [1–3]. It involves partitioning images into distinct regions corresponding to anatomical structures, organs, or pathological features. The increasing number and complexity of medical images have created a tremendous annotation burden for radiologists, calling for an efficient and accurate medical image analysis model for automated or semi-automated segmentation [1].

Inspired by SAM [4], a natural image segmentation foundation model, MedSAM [2] emerges as a pioneering foundation model in the medical field. Different from the traditional task-specific models [5], MedSAM has the potential to tackle the challenge of multimodal medical image segmentation by pre-training on different modalities and different segmentation targets with a box prompt provided by users. The success of foundation models makes it unnecessary to redesign and retrain the network for each task, thereby, simplifying the clinical settings [1]. Based on SAM and MedSAM, many foundation models with more accurate segmentation or more efficient inference schemes were developed [6–11].

Corresponding author: Jiajing Zhang (jiajingz@connect.hku.hk).

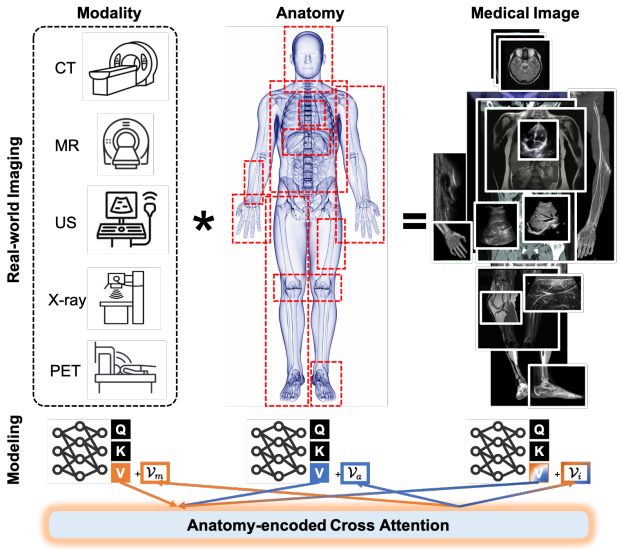

Fig. 1. The convolution relationship of the real-world medical imaging process, which can be modeled by our ASAM.

However, directly transferring models for natural images might not be the best solution for medical images. Firstly, none of the existing foundation models focus on the biggest difference between natural images and medical images — anatomy features. Anatomical features play a pivotal role in medical image segmentation because they provide essential context for distinguishing different structures inside the human body. Anatomy knowledge allows for precise segmentation even in challenging scenarios [12], where structures may overlap or be obscured by artifacts. Secondly, the imaging principles of medical images differ substantially from those of natural images, leading to different sensing and reconstruction paradigms [13, 14]. Natural images are produced using a camera sensing the natural visible light, whereas medical images are formed by sensing the interaction between the biological tissues and specialized external stimuli to portray specific tissue properties and internal tissue structures. For example, in X-ray and computed tomography (CT) imaging, X-ray is generated and attenuated differently through different tis-

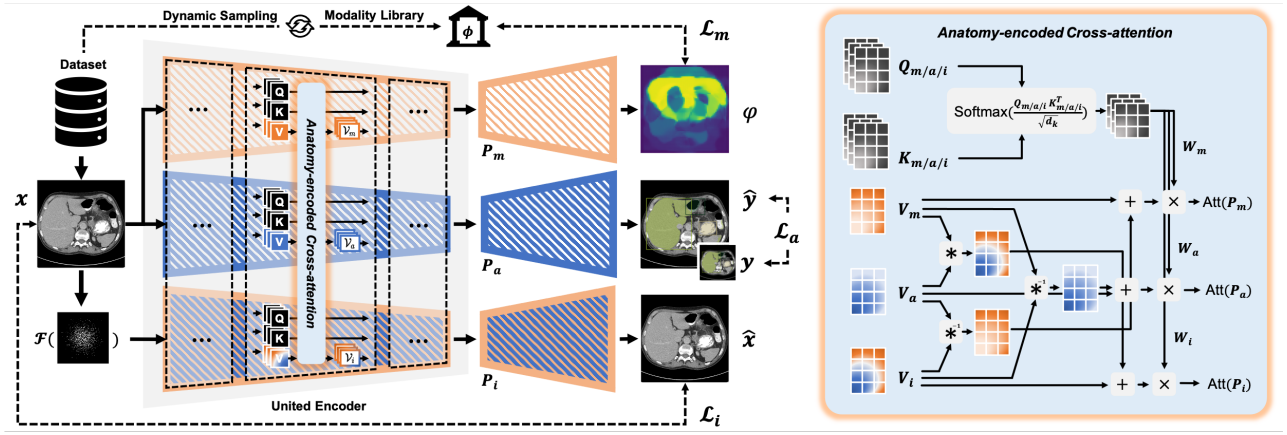

Fig. 2. The model architecture of ASAM, which includes three paths (P_a, P_i, P_m) and unites their encoders into one to perform anatomy-encoded cross attention.

sue layers via ionizing radiation. In ultrasound imaging, MHz sound waves are generated and the echoes returning from different tissues are detected for image formation. Currently, most medical foundation models [2, 6–10, 15] still adopt the processing paradigm designed for natural images and do not consider anatomy features and medical image formation process.

To address the domain gap between natural and medical images, we propose an Anatomy-encoded SAM (ASAM). As shown in Fig. 1, ASAM emulates the real-world medical imaging process by an anatomy-encoded cross-attention and thus enables unsupervised modeling of the anatomy features in multimodal medical images.

2. METHODS

Let $I(i,j)$ denote the observed intensity at pixel (i,j) in a medical image, A represent the true distribution of anatomical structures within the human body, and f_{PSF} be the point spread function (PSF) associated with the imaging modality. The image formation process can be formulated as:

$$\begin{aligned} I(i,j) &= A * f_{PSF} \\ &= \int\int A(u,v) \times f_{PSF}(i-u, j-v)\, du\, dv \end{aligned} \tag{1}$$

Specifically, in CT and X-ray imaging, the PSF is shaped by the radiation beam geometry [16, 17]; in MR imaging, it is determined by radiofrequency (RF) pulse sequences and gradient fields [18], in ultrasound imaging, the PSF is dictated by the transducer and beamformer, with additional effects from sound wave scattering [19]; in PET imaging, the PSF is influenced by the detector characteristics as well as the attenuation and scattering of gamma rays [20]. Thus, the medical image formation process can be characterized as a convolution of modality-specific features with the tissue properties of the individual human body.

Based on this fundamental principle of imaging physics, a medical image can be decoded into modality features and anatomical features. As Fig. 2, ASAM employs three paths to model full image, anatomical features, and modality features, respectively. Each attention operation across these paths interacts through anatomy-encoded cross-attention, simulating the imaging process and facilitating the unsupervised extraction of anatomical features.

2.1. Modeling Imaging Process

Firstly, medical image segmentation essentially identifies a tissue or lesion and outlines its anatomical structure. Therefore, ASAM's anatomy path performs image segmentation tasks to model anatomy features: $\widehat{y} = P_a(x)$, where $x \in \mathbb{R}^{H\times W\times C}$ denotes the input image and $\widehat{y} \in \mathbb{R}^{H\times W\times 1}$ denotes the segmentation map. Following the workflow of MedSAM, this segmentation is prompted by a bounding box.

Secondly, image reconstruction requires both modality and anatomy information. Therefore, ASAM's full-image path performs a self-supervised image reconstruction task to model full-image features. Before entering this path, input images have 95% of their frequency components masked by Gaussian random sampling (M_g). The full-image path aims to reconstruct those severely degraded images to the original high-resolution images: $\widehat{x} = P_i(\mathcal{F}^{-1}(\mathcal{F}(x) \times M_g))$, where $\widehat{x} \in \mathbb{R}^{H\times W\times C}$ denotes the reconstructed results, and $\mathcal{F}(\cdot)$ means the Fourier transform.

Medical images within the same modality inherently exhibit similar modality features. For each modality, we construct a dynamic library (ϕ), which randomly selects five images from the dataset (Φ) belonging to the respective modality while training the model. Hence, $\phi = \{\widetilde{x}_k \mid k = 1 \sim 5\}$, $\widetilde{x}_k \in \Phi$. In this way, a modality library contains unstable anatomical structures but consistent modality features. The modality path aims to decode the modality map from the med-

ical image: $\varphi = P_m(x)$, where $\varphi \in \mathbb{R}^{H\times W\times 1} \mapsto \phi$ should converge to modality commonalities within that library.

ASAM's three paths are all constructed by TinyViT [21] and initialized with the pre-trained LiteMedSAM [2].

2.2. Anatomy-encoded Cross-attention

P_a, P_i, and P_m share the same model structure, making every attention operation in three paths a one-to-one correspondence. ASAM unites the encoders from the three paths into one encoder. Therein, three self-attentions at the same step are reorganized into an anatomy-encoded cross-attention that complies with the imaging process.

Let $\{Q_a, K_a, V_a\}$, $\{Q_i, K_i, V_i\}$, and $\{Q_m, K_m, V_m\}$ be the queries, keys, and values of self-attentions of P_a, P_i, and P_m at a given step, respectively. Based on the convolution relationship in the imaging process, anatomy-encoded cross-attention associates V_i, V_a, and V_m with each other according to Eq. 2∼4 and generates new $\mathcal{V}_i$, $\mathcal{V}_a$, and $\mathcal{V}_m$:

$$\mathcal{V}_i = V_m * V_a = \mathcal{F}^{-1}\left[\mathcal{F}(V_m) \times \mathcal{F}(V_a)\right], \tag{2}$$

$$\mathcal{V}_a = V_i * / * V_m = \mathcal{F}^{-1}\left[\frac{\mathcal{F}(V_i)}{\mathcal{F}(V_m)}\right], \tag{3}$$

$$\mathcal{V}_m = V_i * / * V_a = \mathcal{F}^{-1}\left[\frac{\mathcal{F}(V_i)}{\mathcal{F}(V_a)}\right], \tag{4}$$

where new full-image feature $\mathcal{V}_i$ is the convolution of anatomy feature V_a and modality feature V_m; $\mathcal{V}_a$ is the deconvolution of V_i and V_m; similarly, $\mathcal{V}_m$ is the deconvolution of V_i and V_a. For computational efficiency, ASAM calculates the convolution and deconvolution in the Fourier domain.

In ASAM's united encoder, for each attention step, $P_{a/i/m}$ firstly generates original $\{Q_{a/i/m}, K_{a/i/m}, V_{a/i/m}\}$. The three value vectors mimic the real-world imaging process to yield $\mathcal{V}_{a/i/m}$. Therefore, our cross-attention output is

$$\text{Cross-attention}\begin{bmatrix} Q_i & K_i & V_i \\ Q_a & K_a & V_a \\ Q_m & K_m & V_m \end{bmatrix} = \begin{bmatrix} \text{softmax}\left(Q_i K_i^T/\sqrt{d_k}\right) \\ \text{softmax}\left(Q_a K_a^T/\sqrt{d_k}\right) \\ \text{softmax}\left(Q_m K_m^T/\sqrt{d_k}\right) \end{bmatrix} \times \begin{bmatrix} V_i + \mathcal{V}_i \\ V_a + \mathcal{V}_a \\ V_m + \mathcal{V}_m \end{bmatrix}^T, \tag{5}$$

where $\mathcal{V}_{a/i/m}$ and $V_{a/i/m}$ are summed as residual connection. All self-attention in the encoder of the original TinyViT is replaced by ASAM's cross-attention. In contrast, the decoder part remains unchanged, thereby keeping three paths independent to perform different tasks. This structure allows ASAM to extract anatomy features from the input medical images based on the imaging process via our united encoder, and further segments the anatomical organs/lesions via the decoder of P_a.

As formulated in Eq. 6, our loss function includes unweighted anatomy loss ($\mathcal{L}_a$), image loss ($\mathcal{L}_i$), and modality loss($\mathcal{L}_m$), which are implemented by a Dice loss, a binary cross-entropy loss, and two mean squared error losses.

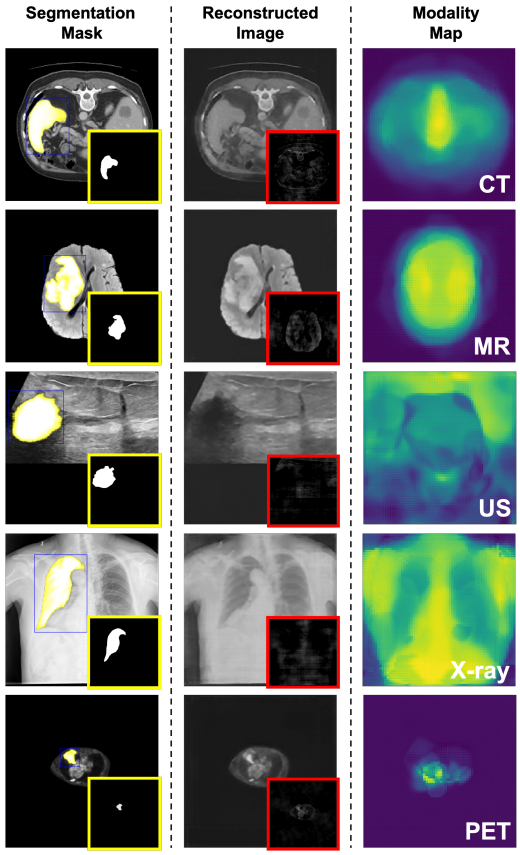

Fig. 3. The predictions of ASAM in five modalities, where the yellow boxes show the segmentation labels and the red boxes show the degraded inputs.

$$\begin{aligned}\mathcal{L}_{ASAM} =& \mathcal{L}_a + \mathcal{L}_i + \mathcal{L}_m = [\mathcal{L}_{\text{Dice}}(\widehat{y}, y) + \mathcal{L}_{\text{BCE}}(\widehat{y}, y)] \\ &+ \mathcal{L}_{\text{MSE}}(\widehat{x}, x) + \frac{1}{|\phi|}\sum_{\widetilde{x}_k \in \phi} \mathcal{L}_{\text{MSE}}(\varphi, \widetilde{x}_k),\end{aligned} \tag{6}$$

where $\widehat{y}$ denotes the predicted segmentation mask and y the denotes label; $\widehat{x}$ denotes the reconstructed image and x denotes the original ground truth; φ denotes the modality map and $\widetilde{x}_k$ denotes an image in the dynamic modality library (ϕ). The labels of $\mathcal{L}_i$ and $\mathcal{L}_m$ come from the dataset itself, so P_i and P_m are unsupervised.

3. EXPERIMENTS AND RESULTS

3.1. Dataset and Implementation Details

We use the multimodal public dataset from the challenge of Segment Anything in Medical Images on Laptop in CVPR 2024 [22], where 17 CT, 11 MR, 3 US, 5 X-ray, and 1 PET datasets are extracted. This large-scale dataset covers

Table 1. Evaluation Metrics of ASAM on five Modalities

Modality		CT	MR	US	X-ray	PET	Avg.
Pre-train	Dice (%)	90.28	87.63	96.60	96.93	85.86	**91.46**
	IoU (%)	82.67	78.74	93.54	94.14	76.87	**85.27**
	HD95 (↓)	15.09	11.49	6.16	6.57	9.30	**9.72**
Fine-tune	Dice (%)	90.60	90.70	97.86	97.19	90.08	**93.29**
	IoU (%)	83.27	83.07	95.85	94.55	82.04	**87.67**
	HD95 (↓)	14.07	10.32	3.73	5.51	4.23	**7.57**

whole-body anatomical targets, including abdomen, brain, head-neck, prostate, heart, chest, lung, kidney, liver, pelvic, breast, and various blood vessels. Both normal organs and lesions are labeled. Such a comprehensive dataset makes ASAM feasible to learn robust anatomy features across different modalities.

ASAM is built with Python 3.10, Pytorch 2.1.2, and CUDA 11.8. One NVIDIA A100 SXM4 and two NVIDIA RTX 3090 GPUs were used for training. In parallel, the modality library was dynamically updated on an AMD EPYC 7532 CPU. We used the AdamW optimizer with L2 regularization for training.

3.2. Experimental Results

We first pre-trained ASAM on the full multimodal dataset for 100 epochs and then fine-tuned it for each modality for another 50 epochs. Dice score, intersection over union (IoU), and Hausdorff distance at 95% (HD95) were evaluated for the pre-trained and fine-tuned models on all five modalities. The three metric values are summarized in Table 1. In the pre-train phase, ASAM achieved a high average Dice score of 91.46%, average IoU of 48.04%, and average HD95 of 9.23. Moreover, ASAM's performance was further improved in the fine-tuning stage, indicating its potential as a foundation model.

Fig. 3 exemplifies predictions of ASAM for five modalities. (1) The first column shows the segmentation masks, where the probability maps from P_a are overlaid on the images, showing ASAM's strong capability to extract fine and complex boundaries signifying anatomical structure. The probability distribution in our predicted mask follows the anatomical textures in the gray-scale image, especially for the MR (second row) and X-ray (fourth row) cases. (2) The second column is the reconstructed image from P_i. Through our cross-attention mechanism, P_i incorporates both anatomy and modality information from P_a and P_m in every attention operation. Fusion of imaging process-based compulsory features allows ASAM to reconstruct a full image matching the correct modality and anatomy from only 5% frequency data. (3) In our united encoder, ASAM decodes the modality features in each attentional operation by deconvolution, and then P_m's decoder converts them into a modality map (the third column). The dynamic update of modality libraries allows P_m to learn to output a modality map that approximates general modality characteristics.

Table 2. Dice Score of ASAM and 10 Comparison Models

Modality	CT	MR	US	X-ray	PET	Avg.
nnUNet [23]	90.05	82.22	80.85	79.54	45.84	**75.70**
MedSAM [2]	88.31	<u>92.92</u>	95.35	91.20	88.75	**91.31**
MedSAM2 [15]	91.67	88.63	78.73	78.22	77.27	**82.90**
LiteMedSAM [21]	92.26	89.63	94.77	75.83	51.58	**80.81**
Swin-LiteMedSAM [6]	91.46	87.12	85.57	83.98	69.43	**83.51**
MedficientSAM [7]	92.15	86.98	82.50	80.47	73.00	**83.02**
Class-prompt MedSAM [8]	92.26	89.63	94.77	76.74	70.28	**84.74**
DAFT [9]	<u>93.14</u>	88.21	94.77	77.07	71.46	**84.93**
OneSAM [10]	89.92	83.28	94.78	75.83	65.92	**81.95**
Gray's Anatomy MedSAM [11]	84.80	82.69	95.01	95.17	79.15	**87.36**
ASAM	90.28	87.63	<u>97.86</u>	<u>97.19</u>	<u>85.86</u>	**<u>91.46</u>**

We compared ASAM with 10 state-of-the-art foundation models in the field of medical image segmentation, including MedSAM2 [15], MedSAM [2] and its lightweight version LiteMedSAM [21] (ASAM's backbone), Swin-LiteMedSAM [6], MedficientSAM [7], Class-prompt MedSAM [8], DAFT [9], OneSAM [10], Gray's Anatomy MedSAM [11], and a CNN-based model – nnUNet [23]. As summarized in Table 2, ASAM achieved the highest average dice score and a balanced performance among five modalities. Specifically, ASAM ranked first in US and X-ray modalities, and was the runner-up in the PET modality following MedSAM. ASAM is slightly degraded in CT and MR modalities compared to the backbone while significantly boosting performance in the other modalities.

A similar trend was observed in Gray's Anatomy MedSAM [11], which also introduces anatomy features. This trade-off between CT/MR and the other modalities reveals that the high performance of existing models relies on certain modality-specific black-box features rather than an anatomical understanding of medical images. Moreover, the data size of CT/MR was much larger than the combined size of the other modalities. This caused the pre-trained backbone to mainly master CT/MR-specific segmentation methods. When we imposed additional anatomy knowledge on the pre-trained backbone via anatomy-encoded cross-attention, the original CT/MR-specific knowledge was corrupted. In contrast, the model learned to be more anatomically robust for other modalities.

4. CONCLUSION

This study proposes an Anatomy-encoded SAM. Our anatomy-encoded cross-attention enables decoding of anatomy features from medical images for high-quality multimodal segmentation. Thanks to the efficient modeling of full-image, anatomy, and modality features, ASAM has the potential to be a truly foundational model for medical image analysis, for not only segmentation but also image reconstruction and synthesis.

5. ACKNOWLEDGMENTS

This work was supported in part by the Hong Kong Health and Medical Research Fund (08192616).

6. REFERENCES

[1] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[2] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, pp. 654, 2024.

[3] Jiajing Zhang, Wenqing Zhang, Haodong Liu, Yingying Liu, Ningning Chen, Jianjia Zhang, and Changhong Wang, "Automatic centroid angle measurement from ct image for preoperative rod design of robotic-assisted screw-rod system implantation," *IEEE Transactions on Medical Robotics and Bionics*, 2024.

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al., "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[5] Jiajing Zhang, Lin Gu, Guanghui Han, and Xiujian Liu, "Attr2u-net: A fully automated model for mri nasopharyngeal carcinoma segmentation based on spatial attention and residual recurrent convolution," *Frontiers in Oncology*, vol. 11, pp. 816672, 2022.

[6] Ruochen Gao, Donghang Lyu, and Marius Staring, "Swin-litemedsam: A lightweight box-based segment anything model for large-scale medical image datasets," *arXiv preprint arXiv:2409.07172*, 2024.

[7] Bao-Hiep Le, Dang-Khoa Nguyen-Vu, Trong-Hieu Nguyen-Mau, Hai-Dang Nguyen, and Minh-Triet Tran, "Medficientsam: A robust medical segmentation model with optimized inference pipeline for limited clinical settings," in *CVPR 2024: Segment Anything In Medical Images On Laptop*, 2024.

[8] Haotian Guan, Bingze Dai, and Jiajing Zhang, "Lite class-prompt tiny-vit for multi-modality medical image segmentation," in *CVPR 2024: Segment Anything In Medical Images On Laptop*.

[9] Alexander Pfefferle, Lennart Purucker, and Frank Hutter, "Daft: Data-aware fine-tuning of foundation models for efficient and effective medical image segmentation," in *CVPR 2024: Segment Anything In Medical Images On Laptop*, 2024.

[10] Khanh-Binh Nguyen and Chae Jung Park, "Onesam: Modality-agnostic for segment anything model in medical images," in *CVPR 2024: Segment Anything In Medical Images On Laptop*.

[11] YoungHwan Choi, In Kyu Lee, and Jonghoe Ku, "Gray's anatomy for segment anything model: Optimizing grayscale medical images for fast and lightweight segmentation," in *CVPR 2024: Segment Anything In Medical Images On Laptop*.

[12] Xu Chen, Chunfeng Lian, Li Wang, Hannah Deng, Tianshu Kuang, Steve Fung, Jaime Gateno, Pew-Thian Yap, James J Xia, and Dinggang Shen, "Anatomy-regularized representation learning for cross-modality medical image segmentation," *IEEE transactions on medical imaging*, vol. 40, no. 1, pp. 274–285, 2020.

[13] Paul Suetens, *Fundamentals of medical imaging*, Cambridge university press, 2017.

[14] Zhifan Gao, Yifeng Guo, Jiajing Zhang, Tieyong Zeng, and Guang Yang, "Hierarchical perception adversarial learning framework for compressed sensing mri," *IEEE Transactions on Medical Imaging*, vol. 42, no. 6, pp. 1859–1874, 2023.

[15] Jun Ma, Sumin Kim, Feifei Li, Mohammed Baharoon, Reza Asakereh, Hongwei Lyu, and Bo Wang, "Segment anything in medical images and videos: Benchmark and deployment," *arXiv preprint arXiv:2408.03322*, 2024.

[16] Masaki Ohkubo, Shinichi Wada, Satoshi Ida, Masayuki Kunii, Akihiro Kayugawa, Toru Matsumoto, Kanae Nishizawa, and Kohei Murao, "Determination of point spread function in computed tomography accompanied with verification," *Medical physics*, vol. 36, no. 6Part1, pp. 2089–2097, 2009.

[17] Kevin T Joyce, Johnathan M Bardsley, and Aaron Luttman, "Point spread function estimation in x-ray imaging with partially collapsed gibbs sampling," *SIAM Journal on Scientific Computing*, vol. 40, no. 3, pp. B766–B787, 2018.

[18] Matthew D Robson, John C Gore, and R Todd Constable, "Measurement of the point spread function in mri using constant time imaging," *Magnetic resonance in medicine*, vol. 38, no. 5, pp. 733–740, 1997.

[19] Christoph Dalitz, Regina Pohle-Frohlich, and Thorsten Michalk, "Point spread functions and deconvolution of ultrasonic images," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 62, no. 3, pp. 531–544, 2015.

[20] E Rapisarda, V Bettinardi, Kris Thielemans, and MC Gilardi, "Image-based point spread function implementation in a fully 3d osem reconstruction algorithm for pet," *Physics in medicine & biology*, vol. 55, no. 14, pp. 4131, 2010.

[21] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan, "Tinyvit: Fast pretraining distillation for small vision transformers," in *European conference on computer vision (ECCV)*, 2022.

[22] Jun Ma, Feifei Li, Sumin Kim, Reza Asakereh, Bao-Hiep Le, Dang-Khoa Nguyen-Vu, Alexander Pfefferle, Muxin Wei, Ruochen Gao, Donghang Lyu, et al., "Efficient medsams: Segment anything in medical images on laptop," *arXiv preprint arXiv:2412.16085*, 2024.

[23] Raphael Stock, Yannick Kirchhoff, Maximilian Rouven Rokuss, Ashis Ravindran, and Klaus Maier-Hein, "Segment anything in medical images with nnunet," in *CVPR 2024: Segment Anything In Medical Images On Laptop*.