FastSAM3D: An Efficient Segment Anything Model for 3D Volumetric Medical Images

Yiqing Shen[1], Jingxing Li[1], Xinyuan Shao[1], Blanca Inigo Romillo[1], Ankush Jindal[1], David Dreizin[2(✉)], and Mathias Unberath[1(✉)]

[1] Johns Hopkins University, Baltimore, MD, 21218, USA
[2] University of Maryland School of Medicine and R Adams Cowley Shock Trauma Center, Baltimore, MD, 21201, USA
{yshen92,unberath}@jhu.edu, daviddreizin@gmail.com

Abstract. Segment anything models (SAMs) are gaining attention for their zero-shot generalization capability in segmenting objects of unseen classes and in unseen domains when properly prompted. Interactivity is a key strength of SAMs, allowing users to iteratively provide prompts that specify objects of interest to refine outputs. However, to realize the interactive use of SAMs for 3D medical imaging tasks, rapid inference times are necessary. High memory requirements and long processing delays remain constraints that hinder the adoption of SAMs for this purpose. Specifically, while 2D SAMs applied to 3D volumes contend with repetitive computation to process all slices independently, 3D SAMs suffer from an exponential increase in model parameters and FLOPS. To address these challenges, we present `FastSAM3D` which accelerates SAM inference to 8 milliseconds per $128 \times 128 \times 128$ 3D volumetric image on an NVIDIA A100 GPU. This speedup is accomplished through 1) a novel layer-wise progressive distillation scheme that enables knowledge transfer from a complex 12-layer ViT-B to a lightweight 6-layer ViT-Tiny variant encoder without training from scratch; and 2) a novel 3D sparse flash attention to replace vanilla attention operators, substantially reducing memory needs and improving parallelization. Experiments on three diverse datasets reveal that `FastSAM3D` achieves a remarkable speedup of 527.38× compared to 2D SAMs and 8.75× compared to 3D SAMs on the same volumes without significant performance decline. Thus, `FastSAM3D` opens the door for low-cost truly interactive SAM-based 3D medical imaging segmentation with commonly used GPU hardware. Code is available at https://github.com/arcadelab/FastSAM3D.

Keywords: Foundation Model · Segment Anything Model (SAM) · Interactive Segmentation · Model Acceleration.

1 Introduction

In medical image analysis, object segmentation is a key aspect of diagnosis- and prognosis-related tasks including lesion localization, tissue characterization, and volume estimation, among others [8,21,29,14]. Traditionally, deep learning models like U-Net [22] and variants [15,11] have excelled in specific tasks and datasets

Table 1. Comparison of SAM approaches regarding applicability for medical imaging, suitability for 3D volumetric data, and computational efficiency of the core components: image encoder, prompt encoder, and mask decoder. The vanilla `SAM` [16] lacks in all criteria. `MobileSAM` [32] improves encoder efficiency, while `TinySAM` [24] accelerates all components, but neither addresses 3D medical imaging data. `MedSAM` [18] and `SAM-Med2D` [4] are tailored for medical 2D data yet do not improve efficiency. `SAM-Med3D` [27] handles 3D medical data but inference times for these limit or altogether preclude real-time interactive use with standard GPU hardware. Our proposed `FastSAM3D` meets all criteria, providing a comprehensive solution for efficient interactive medical image segmentation in volumetric 3D data.

Method	Medical	Volumetric 3D Data	Efficient Image Encoder	Efficient Prompt Encoder	Efficient Mask Decoder
`SAM` [16]	✗	✗	✗	✗	✗
`MobileSAM` [32]	✗	✗	✓	✗	✗
`TinySAM` [24]	✗	✗	✓	✓	✓
`MedSAM` [18]	✓	✗	✗	✗	✗
`SAM-Med2D` [4]	✓	✗	✗	✗	✗
`SAM-Med3D` [27]	✓	✓	✗	✗	✗
`FastSAM3D` (ours)	✓	✓	✓	✓	✓

with clear and confined scope, but often demonstrate limited generalization. While some work considered interactive segmentation approaches as a means to overcome the limitations of narrowly scoped, task-specific models [1,26], the introduction of Segment Anything Model (`SAM`) [16] initiated a paradigm shift to prompt-based interactive segmentation that now provides competitive performance due to the inherent generalizability of foundation models. `SAM` is comprised of a pre-trained Vision Transformer (ViT) encoder [9], a prompt encoder, and a lightweight decoder that facilitates multi-mask prediction via IoU-based ranking. Trained on over 1 billion masks and 11 million images, `SAM` adapts to new tasks without training [16]. Despite successes on natural images [10,17], direct application of `SAM` to medical segmentation reveals performance gaps compared to task-specific U-Nets [33].

To address this, `MedSAM` [18] and `SAM-Med2D` [4] were tailored for 2D medical data via model fine-tuning. When applied to 3D volumetric data, these approaches under-perform due to slice-wise processing [19,3]. They also suffer from an increased computational cost that is proportional to the number of slices in the volume, as well as the higher input resolution. Addressing this gap, `SAM-Med3D` [27] introduced 3D counterparts of SAM's components and end-to-end 3D training.

Existing medical SAMs also face limitations of long inference times and high computational costs stemming from the Transformer architecture [9]. Prior efforts accelerated 2D SAMs for natural images via approaches such as `FastSAM`, which employs a YOLOv8 as the image encoder [2,35]. However, this CNN-based approach exhibits limitations with small object segmentation and deviates from SAM's interactive prompting design [32,35]. Attempts, more aligned

with the SAM value proposition, such as `MobileSAM`, retain the Transformer encoder while employing distillation to transition from a larger SAM encoder to a more lightweight ViT encoder [32]. Following a similar approach, other works have explored several ViT variants as alternative encoders to balance efficiency and effectiveness [34,25]. `TinySAM` further reduces computational load via post-distillation quantization [24]. However, these advancements remain confined to 2D natural images without delving into efficient volumetric medical segmentation. Table 1 offers a systematic comparison of existing SAMs, highlighting the capabilities of our proposed method in addressing the demands of 3D medical image segmentation with enhanced efficiency across all components.

The contributions of this work are two-fold, summarized as follows: Firstly, we introduce `FastSAM3D`, a markedly more efficient 3D SAM for interactive volumetric medical image segmentation. Rather than costly training from scratch which also leads to the difficulty in convergence [12], we propose a layer-wise progressive distillation approach to transfer representational knowledge from a complex 12-layer ViT-B architecture to an efficient customized 6-layer ViT-Tiny encoder. This retains segmentation performance while significantly enhancing computational efficiency. Secondly, we propose a novel 3D sparse flash attention that replaces the standard self-attention operator in all SAM components, dramatically reducing memory footprint, and enabling parallel processing. Together, these innovations address the efficiency limitations that hinder the implementation of medical SAMs for real-time prompt-based interactive 3D segmentation.

2 Methods

2.1 Architecture Overview of `FastSAM3D`

We introduce `FastSAM3D`, a computationally efficient adaptation of `SAM-Med3D` [27], also designed specifically for efficient interactive 3D medical image segmentation. Adhering to the standard SAM paradigm [16], `FastSAM3D` is comprised of three key modules (Fig. 1): (i) a ViT-based image encoder [9] to obtain volumetric embeddings; (ii) a prompt encoder; and (iii) a mask decoder to project representations back to the segmentation mask. To achieve faster inference, `FastSAM3D` distills knowledge from a high-powered 12-layer ViT-B encoder to a streamlined 6-layer ViT-Tiny variant, substantially reducing computational complexity during encoding. Specifically, aside from having fewer layers, each Transformer block contains only 6 attention heads, in contrast to 12 heads per block in `SAM-Med3D`'s ViT architecture. Moreover, we retain the feed-forward network (FFN) within the first two transformer blocks and omit attention operations [9], incurring minimal impact on performance while amplifying speed [31]. This design choice further contributes to shorter training by requiring fewer layers to align with the teacher during our progressive distillation process.

2.2 Layer-wise Progressive Distillation for the Image Encoder

As the image encoder accounts for a major portion of SAM's computational load, our first focus is transferring knowledge from the heavy ViT-B architecture to

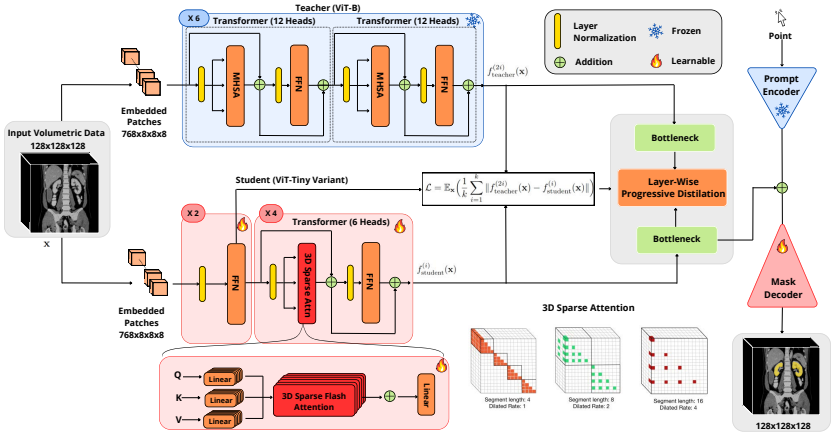

Fig. 1. The overall framework of `FastSAM3D`, comprising a 6-layer ViT-Tiny variant image encoder distilled from a capable 12-layer ViT-B teacher encoder, a lightweight prompt encoder, and a mask decoder. All the self attention operators are replaced by the proposed 3D sparse attention for better efficiency.

a lightweight ViT-Tiny model for efficiency gains. To avoid costly training from scratch, we follow the teacher-student distillation paradigm [12] by designating the 12-layer ViT-B as the teacher model, f_{teacher}, and the 6-layer ViT-Tiny variant as the student model, f_{student}. Unlike traditional logit-level distillation [12] with which all our experiments failed to converge, we propose a novel layer-wise progressive distillation method. This approach allows for a more granular and effective knowledge transfer between the student and teachers by matching the intermediate representation progressively across layers, thus making it easier for optimization. Formally, let $f^{(i)}_{\text{teacher}}(\mathbf{x})$ and $f^{(j)}_{\text{student}}(\mathbf{x})$ denote layer $i = 1, \cdots, 12$ and $j = 1, \cdots, 6$ outputs for an input $\mathbf{x} \in \mathbb{R}^{128\times128\times128}$ from the 12-layer teacher and 6-layer student respectively. The objective of our layer-wise progressive distillation becomes:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}}\Big(\frac{1}{k}\sum_{i=1}^{k} \| f^{(2i)}_{\text{teacher}}(\mathbf{x}) - f^{(i)}_{\text{student}}(\mathbf{x}) \|\Big), \tag{1}$$

where $\|\cdot\|$ denotes the L2-norm, k varies from 1 to 6 based on current and total training iterations:

$$k = \lceil \frac{\#(\text{Current Iteration}) \times 6}{\#(\text{Total Iterations})} \rceil, \tag{2}$$

where $\lceil\cdot\rceil$ is the upper rounding operator. This enables progressive alignment of student and teacher intermediate representations. After finishing layer-wise distillation, we perform logit-level distillation to fit predictions further. In Eq. 1, $\mathbb{E}(\cdot)$ represents the expectation over all possible images.

2.3 3D Sparse Flash Attention

As we observe that the attention operators take up the largest proportion of computation, we introduce a 3D sparse flash attention operator to further enhance efficiency. Specifically, our 3D sparse flash attention scheme supplants the traditional self-attention operation in both the encoder and decoder, integrating extended receptive fields inspired by dilated convolutions [30] with the computational agility achieved by flash attention [6,5].

3D Sparse Attention The 3D sparse attention mechanism aims to expand the receptive field across volumetric data while effectively managing the computational load. Traditional attention mechanisms tend to escalate in computational demand proportional to the increase in data volume, particularly challenging for 3D volumetric data due to its large number of tokens [9,23]. To address this, our approach segments the input token sequence into equally sized partitions of w and applies a strategic sparsification across these segments [7]. This involves selectively sampling data points at the determined intervals with length r, thereby diminishing the overall number of tokens subjected to the attention process. This allows for more efficient computation by focusing attention on fewer yet representative tokens. Formally, the 3D sparse attention mechanism can be formulated as computing the attention over each segment as follows:

$$\hat{S}_i = [S_i, S_{i+r}, \ldots, S_{i+(w-1)r}], \tag{3}$$

where $\hat{S}_i$ represents the selectively sampled segment, ensuring that the model's attention is distributed across a sparse set of points, thereby reducing computational demands without sacrificing the depth of contextual analysis.

Enhancing Efficiency through Parallel Processing with Flash Attention We enhance efficiency by processing each segment in 3D sparse attention independently, enabling parallel operations that significantly boost computational throughput. By incorporating flash attention [6,5], our model optimizes the functionality of parallel attention heads, substantially reducing the time and memory overhead associated with simultaneous processing activities.

Overall Processing Procedure The 3D sparse flash attention operator, integral to both the image encoder and mask decoder, operates through a sequence of orchestrated steps as follows. The process starts with the sparsification step, wherein the input sequence undergoes partitioning into sparse segments. Subsequently, the attention operation ensues, wherein the previous segments are subjected to the flash attention for parallelization [6,5]. The focus of this stage is on harnessing the reduced sequential computation and memory optimization capabilities of flash attention. The final phase is recomposition [7], where the discrete outputs procured from the flash attention are reassembled to form the final encoded representation. This stage ensures that the final encoded representation has an identical dimension to its input.

Table 2. Performance comparison of 2D and 3D SAM approaches in terms of Dice score. We measure the performance at 1, 3, 5, and 10 point prompts (pt). SAM-Med3D and our FastSAM3D are evaluated in a 3D context, whereas SAM, MobileSAM, TinySAM, MedSAM and SAM-Med2D are applied independently to all 2D slices of the entire 3D volume. Notably, FastSAM3D demonstrates competitive performance with SAM-Med3D and shows enhanced Dice scores relative to all its 2D counterparts, highlighting the effectiveness of our approach. The best performance is shown in **red** and boldface, while the second best is in blue.

Dim	Method	AMOS [13]				TotalSegmentator [28]				BraTS [20]			
		1pt	3pt	5pt	10pt	1pt	3pt	5pt	10pt	1pt	3pt	5pt	10pt
2D	SAM [16]	0.049	0.093	0.114	0.145	0.202	0.279	0.311	0.348	0.108	0.192	0.217	0.237
	MobileSAM [32]	0.041	0.056	0.063	0.070	0.149	0.170	0.182	0.212	0.079	0.132	0.156	0.186
	TinySAM [24]	0.049	0.077	0.089	0.101	0.171	0.225	0.243	0.262	0.103	0.165	0.187	0.211
	MedSAM [18]	0.004	0.051	0.060	0.074	0.006	0.069	0.090	0.111	0.008	0.059	0.064	0.071
	SAM-Med2D [4]	0.097	0.127	0.129	0.132	0.008	0.081	0.100	0.128	0.013	0.076	0.082	0.084
3D	SAM-Med3D [27]	**0.289**	**0.386**	**0.418**	**0.448**	**0.252**	**0.400**	**0.463**	**0.522**	0.328	0.395	0.418	**0.446**
	FastSAM3D	0.273	0.368	0.402	0.437	0.250	0.378	0.445	0.519	**0.333**	**0.401**	**0.421**	0.445

3 Experiments

Implementation Details Our method as well as all baseline methods are implemented in Python 3.9 and PyTorch 2.1.0. The computational environment for our experiments is standardized across all methods, utilizing an NVIDIA A100 GPU with 40Gb of memory. For the layer-wise progressive distillation, we set the total training iteration number in Eq. (2) to 36. Training is facilitated by the Adam optimizer, with a learning rate of 5×10^{-3} and a batch size of 16. For evaluation metrics, we use the Dice score to measure segmentation performance. We also report inference time, floating point operations (FLOPs), and memory cost to quantify the computational complexity.

Datasets Our evaluation incorporates three diverse datasets that span two modalities, namely computed tomography (CT) and magnetic resonance imaging (MRI), where we follow the dataset splits of previous work [27]. (1) The *AMOS* dataset [13] is a substantial and varied clinical collection designed for abdominal organ segmentation with 500 CT and 100 MRI scans. (2) The *TotalSegmentator* dataset [28] consists of 1228 CT studies each with 117 anatomical structures acquired from different pathologies, scanners, series, and institutions. (3) The *BraTS 2021* dataset [20] assembles a total number of 1251 multi-institutional MRI scans.

Performance Comparison Table 2 compares segmentation performance for FastSAM3D with various 2D and 3D SAM approaches. Fig. 2 provides an il-

Table 3. Comparison of the computational efficiency with respect to the encoder and decoder. We report the time (ms), FLOPs (G), and memory (Gb), alongside acceleration factors relative to 2D `SAM` [16] and 3D `SAM-Med3D` [27]. For 2D SAMs, we compute the time to process all the slices within volumetric data. The best results are highlighted in **bold** if statistically different from the second best result ($p < 0.01$).

Dim	Method	Resolution	Encoder			Decoder			Acceleration	
			Time (ms)↓	FLOPs (G)↓	Memory (Gb)↓	Time (ms)↓	FLOPs (G)↓	Memory (Gb)↓	To 2D ↑	To 3D ↑
2D	`SAM` [16]	1024 × 1024	3980	369.0	7.87	239	3.0	5.57	1.00×	/
	`MobileSAM` [32]	1024 × 1024	584	36.7	5.48	233	3.0	5.27	5.16×	/
	`TinySAM` [24]	1024 × 1024	609	36.7	5.48	246	3.0	5.27	4.93×	/
	`MedSAM` [18]	1024 × 1024	3983	369.0	7.87	241	2.9	5.57	1.00×	/
	`SAM-Med2D` [4]	256 × 256	1063	32.0	6.32	216	**0.21**	5.55	3.30×	/
3D	`SAM-Med3D` [27]	128 × 128 × 128	70	89.5	6.58	20	2.8	5.53	60.27×	1.00×
	`FastSAM3D`	128 × 128 × 128	**3**	**21.9**	**0.78**	**5**	2.8	**0.71**	**527.38×**	**8.75×**

lustrative visualization for samples segmented by different methods. `FastSAM3D` not only demonstrates competitive segmentation performance in comparison to its teacher model, `SAM-Med3D`, but also surpasses all 2D efficient SAM models, especially when the number of point prompts is increased. For example, `FastSAM3D` achieves a Dice score of 0.437 on the *AMOS* dataset with 10 point prompts, which is a significant improvement over the 0.306 score from the best-performing 2D model ($p < 0.01$). This trend is consistent across the *TotalSegmentator* and *BraTS* datasets, underscoring the robustness of `FastSAM3D` across different datasets, modalities, and organs. Additionally, 2D SAM methods require intensive per-slice prompting as opposed to `FastSAM3D` which only involves volume-level interactions.

Computational Efficiency Comparison Regarding the computational efficiency, Table 3 reveals that `FastSAM3D` reduces the inference time for the encoder to 3 milliseconds and decoder to 5 milliseconds for 3D volumetric images, a substantial improvement from the 3980 milliseconds required by the vanilla `SAM` employed in a slice-by-slice manner. Moreover, `FastMed3D` requires fewer

Table 4. Ablation study for the contribution of 3D sparse attention ('Sparse Attn.') and flash attention ('Flash Attn.') to the performance and efficiency of `FastSAM3D`. Best scores are highlighted in **bold**, if statistically different from the second best result ($p < 0.01$). 3D sparse attention and flash attention contribute to substantial improvements in time and memory requirements without statistically significant performance decline.

Sparse Attn.	Flash Attn.	AMOS [13]				TotalSegmentator [28]				BraTS [20]				Encoder		
		1pt	3pt	5pt	10pt	1pt	3pt	5pt	10pt	1pt	3pt	5pt	10pt	Time	FLOPs	Memory
✗	✗	0.282	0.375	0.403	0.436	0.243	0.371	0.442	0.516	0.335	0.404	0.422	0.444	10	23.1	1.16
✗	✓	0.276	0.366	0.398	0.432	0.247	0.374	0.438	0.516	0.331	0.402	0.421	0.445	6	**21.9**	1.15
✓	✗	0.277	0.370	0.402	0.433	0.255	0.381	0.450	0.520	0.328	0.403	0.422	0.445	9	23.1	0.79
✓	✓	0.273	0.368	0.402	0.437	0.250	0.378	0.445	0.519	0.333	0.401	0.421	0.445	**3**	**21.9**	**0.78**

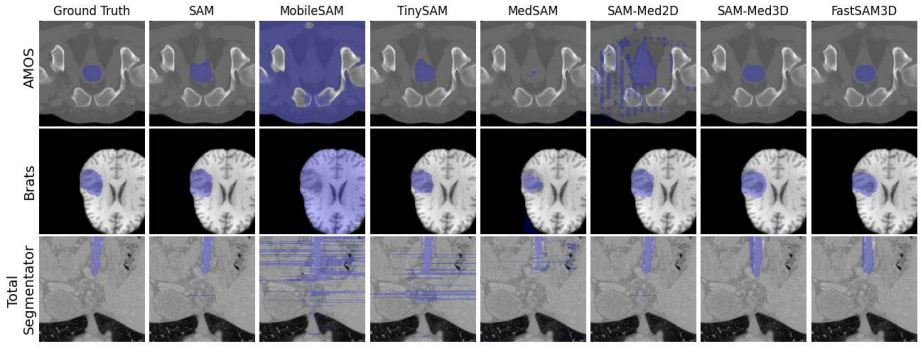

Fig. 2. Representative segmentation results from all methods across three datasets. `FastSAM3D` demonstrates accurate contour delineation comparable to `SAM-Med3D`.

FLOPs and less memory than all counterparts, achieving a 527.38× acceleration compared to the vanilla `SAM` and 8.75× acceleration compared to `SAM-Med3D`.

Ablation Study Table 4 illustrates the effectiveness of both sparse and flash attention in optimizing computational efficiency. Specifically, when neither 3D sparse nor flash attention mechanisms were applied, the model achieved a Dice score of 0.436 on the AMOS dataset with 10 prompts. The introduction of 3D sparse attention marginally reduces the Dice score to 0.433 from 0.436 on *AMOS* but substantially reduces memory consumption from 1.06 Gb to 0.79 Gb. Flash attention alone improves inference time from 10 ms to 6 ms, underscoring its impact on computational efficiency. Moreover, the concurrent implementation of both sparse and flash attention yields the most substantial improvements. For instance, the Dice score on the AMOS dataset with 10 prompts increases to 0.437, and the encoder time is reduced to 3 ms, Memory requirements are also minimized to 0.78 Gb, suggesting an optimized model footprint.

4 Conclusion

We present `FastSAM3D`, an innovative adaptation of `SAM` for efficient segmentation of volumetric medical imaging data. This model addresses the critical challenges of high inference time and the substantial computational cost associated with previous 3D `SAM` methods. Through a novel layer-wise progressive distillation and 3D sparse flash attention integration, we significantly reduce computational demands while maintaining high segmentation performance. Our experiments across different modalities and organs demonstrate that `FastSAM3D` not only accelerates inference by factors of 527.38× compared to 2D `SAM`s and 8.75× to 3D `SAM`s but also retains the flexibility of SAM's interactivity, making it a promising and powerful tool for clinical deployment. `FastSAM3D` opens up the possibility of real-time human-machine interaction by facilitating rapid

prompting volumetric segmentation, thereby potentially maximizing user agency and trust while minimizing effort, workload, and wait time-related frustration. With the speed and efficiency of `FastSAM3D`, another possible direction includes the development of mixed reality (MR) applications for surgical planning and guidance.

Acknowledgements This work was supported in part by grants from the National Institutes of Health (NIH R01 GM148987-01).

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

References

1. Amrehn, M., Gaube, S., Unberath, M., Schebesch, F., Horz, T., Strumia, M., Steidl, S., Kowarschik, M., Maier, A.: Ui-net: Interactive artificial neural networks for iterative image segmentation based on a user model. arXiv preprint arXiv:1709.03450 (2017)
2. Bolya, D., Zhou, C., Xiao, F., et al.: Yolact: Real-time instance segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9157–9166 (2019)
3. Bui, N.T., Hoang, D.H., Tran, M.T., et al.: Sam3d: Segment anything model in volumetric medical images. arXiv preprint arXiv:2309.03493 (2023)
4. Cheng, J., Ye, J., Deng, Z., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
5. Dao, T.: Flashattention-2: Faster attention with better parallelism and work partitioning. arXiv preprint arXiv:2307.08691 (2023)
6. Dao, T., Fu, D., Ermon, S., et al.: Flashattention: Fast and memory-efficient exact attention with io-awareness. Advances in Neural Information Processing Systems **35**, 16344–16359 (2022)
7. Ding, J., Ma, S., Dong, L., et al.: Longnet: Scaling transformers to 1,000,000,000 tokens. arXiv preprint arXiv:2307.02486 (2023)
8. Dora, L., Agrawal, S., Panda, R., et al.: State-of-the-art methods for brain tissue segmentation: A review. IEEE reviews in biomedical engineering **10**, 235–249 (2017)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Gao, H., Li, Y., Long, K., et al.: A survey for foundation models in autonomous driving. arXiv preprint arXiv:2402.01105 (2024)
11. He, Z., Unberath, M., Ke, J., et al.: Transnuseg: A lightweight multi-task transformer for nuclei segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 206–215. Springer (2023)
12. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
13. Ji, Y., Bai, H., Ge, C., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in Neural Information Processing Systems **35**, 36722–36732 (2022)

14. Jiang, H., Diao, Z., Yao, Y.D.: Deep learning techniques for tumor segmentation: a review. The Journal of Supercomputing **78**(2), 1807–1851 (2022)
15. Ke, J., Lu, Y., Shen, Y., et al.: Clusterseg: A crowd cluster pinpointed nucleus segmentation framework with cross-modality datasets. Medical Image Analysis **85**, 102758 (2023)
16. Kirillov, A., Mintun, E., Ravi, N., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
17. Liu, X., Zhou, T., Wang, Y., et al.: Towards the unification of generative and discriminative visual foundation model: A survey. arXiv preprint arXiv:2312.10163 (2023)
18. Ma, J., He, Y., Li, F., et al.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
19. Mazurowski, M.A., Dong, H., Gu, H., et al.: Segment anything model for medical image analysis: an experimental study. Medical Image Analysis **89**, 102918 (2023)
20. Menze, B.H., Jakab, A., Bauer, S., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
21. Mirikharaji, Z., Abhishek, K., Bissoto, A., et al.: A survey on deep learning for skin lesion segmentation. Medical Image Analysis p. 102863 (2023)
22. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation pp. 234–241 (2015)
23. Shen, Y., Guo, P., Wu, J.a.: Movit: Memorizing vision transformers for medical image analysis. In: International Workshop on Machine Learning in Medical Imaging. pp. 205–213. Springer (2023)
24. Shu, H., Li, W., Tang, Y., Zhang, Y., Chen, Y., Li, H., Wang, Y., Chen, X.: Tinysam: Pushing the envelope for efficient segment anything model. arXiv preprint arXiv:2312.13789 (2023)
25. Wang, A., Chen, H., Lin, Z., et al.: Repvit-sam: Towards real-time segmenting anything. arXiv preprint arXiv:2312.05760 (2023)
26. Wang, G., Li, W., Zuluaga, M.A., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., et al.: Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE transactions on medical imaging **37**(7), 1562–1573 (2018)
27. Wang, H., Guo, S., Ye, J., et al.: Sam-med3d. arXiv preprint arXiv:2310.15161 (2023)
28. Wasserthal, J., Breit, H.C., Meyer, M.T., et al.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5) (2023)
29. Wu, J., Xu, Q., et al.: Swin transformer improves the idh mutation status prediction of gliomas free of mri-based tumor segmentation. Journal of Clinical Medicine **11**(15), 4625 (2022)
30. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
31. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: Metaformer is actually what you need for vision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10819–10829 (2022)
32. Zhang, C., Han, D., Qiao, Y., et al.: Faster segment anything: Towards lightweight sam for mobile applications. arXiv preprint arXiv:2306.14289 (2023)
33. Zhang, Y., Shen, Z., Jiao, R.: Segment anything model for medical image segmentation: Current applications and future directions. arXiv preprint arXiv:2401.03495 (2024)

34. Zhang, Z., Cai, H., Han, S.: Efficientvit-sam: Accelerated segment anything model without performance loss. arXiv preprint arXiv:2402.05008 (2024)
35. Zhao, X., Ding, W., An, Y., et al.: Fast segment anything. arXiv preprint arXiv:2306.12156 (2023)