UNIVERSITY OF CAPE TOWN

DEPARTMENT OF COMPUTER SCIENCE

CS Honours Project Final Paper 2024

Title:

Fine Tuning SAM Through Low Rank Adaptation (LoRA)

Author: Tapera Chikumbu

Project Abbreviation: SAMSEG

Supervisor(s): Patrick Marais, Fred Nicolls

Category	*Min*	*Max*	Chosen
Requirement Analysis and Design	*0*	*20*	5
Theoretical Analysis	*0*	*25*	0
Experiment Design and Execution	*0*	*20*	20
System Development and Implementation	*0*	*20*	0
Results, Findings and Conclusions	*10*	*20*	20
Aim Formulation and Background Work	*10*	*15*	15
Quality of Paper Writing and Presentation	*10*		10
Quality of Deliverables	*10*		10
Overall General Project Evaluation (*this section allowed only with motivation letter from supervisor*)	*0*	*10*	0
Total marks		**80**	**80**

Fine Tuning SAM Through Low Rank Adaptation (LoRA)

Tapera Chikumbu
University of Cape Town
Cape Town, South Africa

ABSTRACT

Segment Anything Model (SAM) is a transformer based model for natural image segmentation. It was developed to be a foundation model that can easily generalise to tasks it was not specifically trained on. However, SAM's zero-shot performance on tumour segmentation is below that of pre-trained state of the art models. This paper investigates the use of low rank adaptation (LoRA) to efficiently train SAM for tumour segmentation. Experiment results show improved segmentation accuracy after applying LoRA. This represented by a change in Dice-Cross Entropy loss on the validation set from an initial value of 0.1 to a minimum of 0.02 with rank-64 adapter layers fitted to the image encoder. However, this performance increase was fairly inconsistent across ranks and Dice scores could not be reliably calculated for all models. Further experiments are required for a solid conclusion.

KEYWORDS

Image Processing, Computer Vision, Machine learning, Segment Anything Model

1 INTRODUCTION

Semantic image segmentation is the grouping of related image pixels into meaningful regions-of-interest. This creates a detailed understanding of the image scene and the features it contains. Real world applications of semantic segmentation are found in fields like robotics, satellite imaging, medical imaging and retail shopping [13]. This paper focuses on segmentation for tumour detection in medical imaging.

A wide array of of machine learning models are capable of semantic segmentation. Meta's Segment Anything Model (SAM) [8] is one such creation that excels at image segmentation. Its transformer architecture makes the model robust enough to perform well on tasks it did not specifically train for. Several researchers have compared SAM's zero-shot performance on medical data to that of state of the art segmentation models like U-Net and BEAL [1, 11**?**]. The results of this research show SAM unable to replicate its performance from natural image segmentation. A number of medical focused SAM variants have been successfully fine tuned.

There are primarily two ways of fine-tuning. Full fine-tuning involves training all parameters in a model to help it capture new data.. The alternative is parameter efficient fine-tuning (PEFT). Only a subset of all available parameters are trained on the new data. Careful selection of what parameters to train can help improve model accuracy while avoiding the performance costs of full fine-tuning. This paper investigates the effect parameter efficient fine-tuning has on SAM's accuracy when applied to medical image segmentation.

The transformer architecture is common among large language models (LLMs). This led us to investigate methods of PEFT used in LLMs when deciding how to fine-tune SAM. Chief among these was the concept of low rank adaptation (LoRA). Here, a model's pre-trained parameters are kept constant during training. New parameters are instead added and trained in parallel to the existing ones. A large enough set of LoRA parameters is said to be capable of fully capturing the additional context of the training examples. A major part of the research was implementing low rank adaptation on SAM and comparing performance with different LoRA configurations.

This paper aims to investigate how adding low-rank adaptation to SAM affects it's performance on brain tumour segmentation. The coming sections start with an explanation of how SAM is currently used for both medical and natural image segmentation. This is followed by a look into how LoRA is used for efficient fine-tuning. Having presented this background information, we proceed to describe the experimental design and implementation of LoRA used. Using these experiment results, we intend to answer the research question of *"Can adding LoRA layers to a pre-trained SAM improve its intracranial meningioma segmentation performance on brain MRIs, compared to the baseline SAM performance on the same task?"*.

2 BACKGROUND AND RELATED WORK

2.1 Transformers and SAM

Transformers were first proposed by Vaswani et al. [14] for natural language processing. They ditch the recurrent layers of traditional neural networks for a combination of feed-forward and attention layers. This motivated the development of Vision Transformers (ViT) [3] for general image processing tasks. A ViT is a trainable block used to encode images as sets of contextualised embeddings. The richness of these feature representations makes them useful for downstream tasks. Three main transformer configurations were investigated by Dosovitskiy e.t. al: "base" (ViT-B), "large" (ViT-L) and "huge" (ViT-H). Each differs in parameter count which translates to the level of context they can capture. Having more parameters makes it easier for long-range dependencies between features to be extracted. However, this increases computing costs as more calculations are required. A trade-off between accuracy and inference time must be made when selecting a configuration.

SAM is a promptable foundation model for natural image segmentation. It was designed with an emphasis on zero-shot transfer [8]. This entails a model being applied to a task it received no training on while maintaining performance. A key factor making this transfer learning possible is the use of a transformer based masked auto-encoder[4] in SAM. The image embeddings it generates are input along with a segmentation prompt to a mask decoder. A full representation of this flow is given in Figure 2.

Variants of the three ViT configurations suggested by Dosovitskiy et. al were implemented as image encoders. Maintaining the "base", "large", "huge" naming scheme, these were pre-trained on a large dataset for SAM. SAM was accompanied by its large training

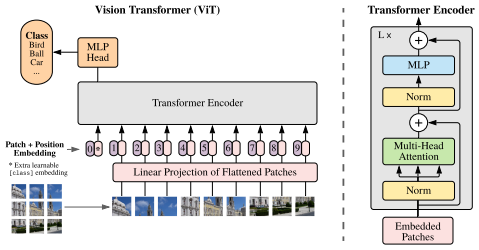

Figure 1: Overview of Vision Transformer showing conversion from 2D image to contextualized embeddings (source)

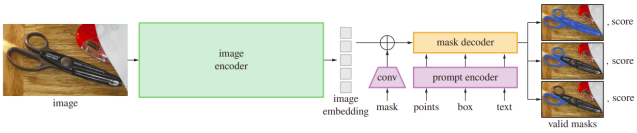

Figure 2: Flow of input through each component of SAM during segmentation. (source)

dataset, SA-1B. Included in the dataset are >1 billion segmentation masks from >11 million high-resolution images [8].

2.2 Medical SAM

Ma et. al [10] performed highly influential research on applying SAM to medical imaging. They developed MedSAM, a foundation model for medical image segmentation. It achieves this using a collection of over a million masked pairs from different anatomical structures and modalities. Having such a diverse training set improves the model's zero-shot transfer. It however requires long pre-training sessions. Several task-specific models have been developed from MedSAM. It is also used as a benchmark for many more models [6, 15].

Wu et. al [15] designed Med-SA to improve on MedSAM. This model introduces a Space-Depth Transpose (SD-Trans) technique. With this, the input spatial dimension is transposed to the depth dimension. Doing so allows the same self-attention blocks to process different dimensional information given different input. As a result, after slicing a 3D image into 2D representations, the information learned from adjacent slices can help improve the segmentation of the current slice.

A different approach was taken by Zhang and Liu [16]. LoRA layers were added to the image encoder. In doing so, pre-trained encoder weights were kept frozen. Fine-tuning instead relied on updates to the prompt encoder, mask decoder and the low rank matrices of the added LoRA layers. The modified SAM was able to reach an impressive Dice score of 81.88 on the Synapse multi-organ segmentation dataset.

2.3 Low-Rank Adaptation (LoRA)

Adapters are compact, trainable layers that can be added to machine learning models. These allow for parameter efficient fine-tuning of pre-trained models by reducing the number of trainable parameters. The standard adapter, as studied by Houlsby et. al. [5], is placed in

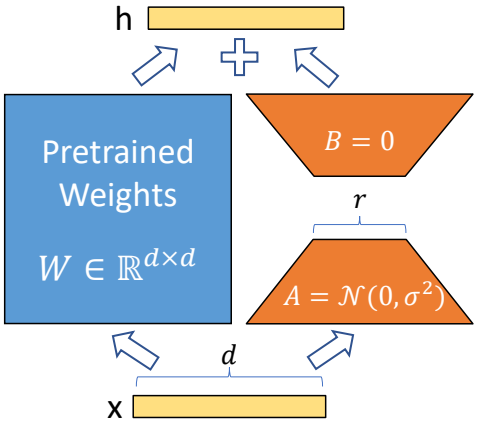

Figure 3: Visualisation of A and B matrices used in LoRA (source)

series with the other blocks of a model. The non-adapter parameters are kept constant (*frozen*) while adapter parameters are sequentially updated as input passes through the model. A downside to this approach is the increased model depth from adding the parameters in series. Every adapter layer added to the model increases the time it takes the to generate predictions from a given input. This is known as inference latency.

LoRA is an alternative approach capable of eliminating this latency . It follows the same procedure of updating the added adapter layers while freezing pre-trained parameters [7]. The key difference lies in adapter placement. LoRA adapters have no inference latency as they are placed parallel to the parameters being fine-tuned.

Rank refers to this shared dimension between an ordered pair of matrices. The full set of pre-trained weights, W, can be expressed as a single d_in x d_out matrix. Low rank approximation decomposes this matrix into an equally representative pair of smaller matrices, A (d_in x rank) and B (rank x d_out) as shown in Figure 3. The dot product (A.B) of the matrices returns a matrix with the same dimensions as the set of pre-trained weights.

Evidence suggests that a reduction in parameters is not always a good thing [7]. Significant computational savings are made possible by only training these smaller A and B matrices. Changing the rank used during decomposition alters the sizes of A and B. A model's ability to capture complex, task-specific patterns is called its representational power. This is often proportional to its parameter count. Only training small sets of parameters therefore limits the complexity of tasks that can be learned. Effective LoRA therefore requires a balance between cost saving and representational power.

3 DESIGN AND IMPLEMENTATION

3.1 Data Selection and Evaluation

The selection of a suitable dataset was required before any experimentation could take place. Factors like the amount of training data, the availability of annotations and the presence of comparable models were used to motivate the selection.

This led to the BraTS 2023 Meningioma Challenge dataset being used for experimentation. It offers a standardised benchmark with

the largest collection of multi-label expert-annotated meningioma mpMRIs to date [9]. A total of 1000 MRIs are provided but only 880 were annotated with ground truth masks. These annotated samples were partitioned as 680 for training and 200 for validation. Each sample contained five distinct images. These consist of four images from different scan types (t1n, t1c, t2f, t2w) with the fifth image representing the ground-truth segmentation mask.

Our focus is on segmentation accuracy. Dice scores will be used to express this accuracy. The Sørensen-Dice coefficient [2] compares the intersection of samples to the sum of the constituent samples and expresses the result as a percentage. In our case, we will be comparing ground-truth masks (y) and the predicted masks ($\hat{y}$) for each model.

$$Dice(y, \hat{y}) = \frac{2|y \cap \hat{y}|}{|y| + |\hat{y}|} \tag{1}$$

$$L_{\text{Dice}}(y, \hat{y}) = 1 - \frac{2|y \cap \hat{y}|}{|y| + |\hat{y}|} \tag{2}$$

Another metric used for training was the binary cross-entropy loss. This makes a pixel level comparison between the model predictions and ground truth outputs

$$L_{\text{CE}}(y, \hat{y}) = \hat{y}[ylog(\hat{y}) + (1\hat{y}y)log(1\hat{y}\hat{y})] \tag{3}$$

A weighted average of the two losses, called the Dice-CE, was used.

$$L_{\text{Dice-CE}} = [w_{CE} * L_{\text{CE}}] + [w_{Dice} * L_{\text{Dice}}] \tag{4}$$

3.2 Preliminary Experiments

A preliminary experimentation phase was used to compare possible approaches for model fine-tuning. This begun with the full fine-tuning of SAM on the chosen dataset. The goal of this was to create a baseline for what performance improvements were possible. The image encoders packaged with SAM do not accept 3D input. Each sample volume was instead horizontally sliced to generate 2D representations as shown in Figure 4. Initial training used every tumour containing slice of the image. Training continued for a total of 150 epochs but the best performance was seen in epoch 125. This is the SAM checkpoint that was used as a baseline for evaluation.

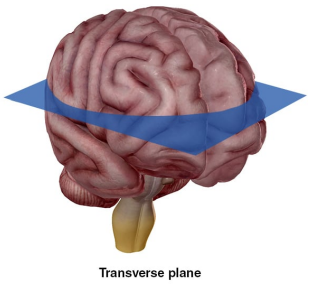

Figure 4: Illustration of the human brain with the transverse anatomical plane highlighted. (source)

Further experiments helped determine the method used when feeding samples to the modified model. Alternative encoders capable of handling the multidimensional nature of the MRIs were looked into [12, 15]. During this step, we attempted to replace SAM's image encoder with the space-depth encoder used in MedSA [15]. This change in architecture proved too complex to implement in the limited time-frame. The alternative approach of slicing the images to create 2D representations was adopted. Slices were taken along the transverse plane as shown in Figure 4.

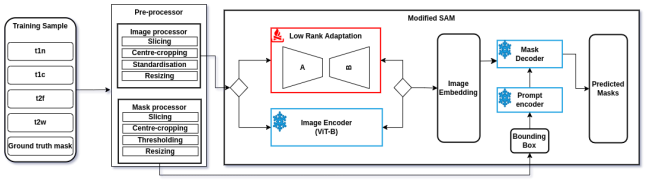

Figure 5: System architecture used during final training. Pre-trained encoder and decoder parameters are kept constant while only the adapter layers are fine-tuned.

An investigation into data batching was also performed. SAM is capable of making simultaneous predictions using batched input images and prompts. Proper utilisation of this feature had the potential to reduce operational overheads and increase model throughput. Hardware constraints however hindered this approach. Only a small amount of the allocated memory was left available during the prediction stage. This meant no meaningful batching could be achieved.

The standardised nature of images across scan types introduces a level of redundancy. Images for the same scan type and those from different scans of the same sample would be highly similar. A comparison of model performance when trained on varying numbers of slices was used to measure these possible redundancies. Four slicing approaches were compared: full volume of slices, every second slice, a random sample of slices and the slice with the largest tumour area (max-slice).

3.3 Final Implementation

Figure 5 illustrates the final model configuration used for experiments. A ViT-B image encoder was used as it was the most lightweight with short training times. This image encoder was modified with LoRA layers being added to every attention block. The rank of these LoRA was kept as a variable. Only the LoRA weights were updated during training with all other model parameters kept constant.

3.3.1 Pre-processing. In line with the preliminary approach, image slices were taken along the transverse plane for each MRI scan type. The different slicing methods investigated all made a trade-off between speed and accuracy. However, the difference in speeds was more pronounced. It was decided to slice the image using the max-slice technique. This offered the fasted inference during model training with only a moderate sacrifice in potential improvements.

Centre cropping was implemented to reduce the amount of border pixels in each dimension. This removes irrelevant data while maintaining image quality. The desired dimensions were found by concatenating all the training images and generating a bounding box around the combined images. It was reasoned that the dimensions of this box would would equal the minimum dimensions

required to fully capture the brain in any given image. Through this process, we found a cropped size of 138 x 202 to be the most appropriate.

The pixel values in each image were also re-scaled to fit a standard normal distribution. This was adopted following best-practices in machine learning. By normalising the input, the total variance is reduced. A result of this is outliers in the data, like high-intensity tumour pixels, become easier to identify.

Further image resizing was necessary before any embeddings could be generated. The image encoder uses a standardised input where longest side of the input must be exactly 1024 pixels. Each slice was thus up-sampled to be 700 x 1024.

A degree of pre-processing was required for the ground truth masks as well. The same slicing, resizing and centre cropping operations used on the image scans were applied to the masks. However, pixel standardisation was replaced by thresholding. This generated binary masks with each pixel marked as either a tumour (1) or non-tumour (0). Bounding box prompts for mask prediction were also generated during this pre-processing step.

3.3.2 Model Training. The model processed the slices from the different scan type sequentially. Each pre-processed slice was passed to the modified LoRA-encoder and used to generate image embeddings. The box prompts from the ground truth masks were then converted into embeddings within the prompt encoder. These two embeddings were used by the mask decoder to generate predictions for each slice. Separate predictions from the different scan types were combined to form an aggregate prediction at the index of the slice being considered. This combined mask was compared to the ground truth mask for optimization.

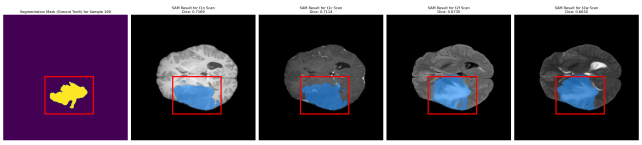

Figure 6: Bounding boxes with generated segmentation masks across the different scan types for one image

The Adam optimizer was used to train the modified SAM. Adam uses the gradient of a loss function with respect to the parameters being optimized to determine how much to update each parameter by. To train the LoRA weights, the optimizer was fit using the image encoder parameters. Equation 4 served as the loss function.

3.3.3 Validation. Model performance was evaluated using the separate validation set. During validation, predictions were made for every tumour containing slice of the image. The prediction quality calculated according to Equation 4

The LoRA adapter used are referenced as LoRA_Rank. For example, LoRA_64 is a LoRA layer with rank 64.

4 RESULTS AND DISCUSSION

The first experiment performed using the modified SAM was to investigate the effect of rank on the initial performance of the model. This was done by comparing the training loss when LoRA layers of ranks 64, 128 and 256 were added. Figure 7 shows the largest performance change was from rank 128 adapters. Losses reduced from 0.3 to 0.17. However, the lowest average losses came from LoRA_64 with a maximum of 0.14 and minimum of 0.06. The loss

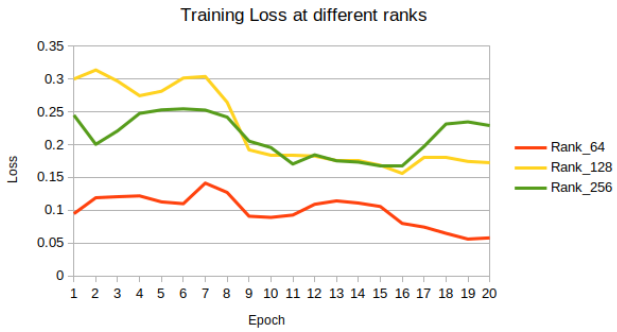

Figure 7: Plot showing change in training loss with different ranks of LoRA adapters

values observed in LoRA_64 and LoRA_256 display unexpected patterns. Training loss is expected to reduce over time. Though there are slight drops below the initial loss value, these configurations display high levels of stagnation. These are periods where the loss remains relatively constant over time. The most prominent of these is in-between epochs 4 and 8. During this period, both configurations have training losses higher than their initial losses. This could be the result of the loss function descending into a local minima. However, similar results could be observed when training parameters are incorrectly configured.

Figure 8 takes a deeper look at the performance of LoRA_64. The model was trained over 50 epochs with validation performance being calculated calculated every two epochs. The first 33 epochs display very positive results. Despite some possible spikes from noise in the data, a general decrease from 0.3 to 0.05 is observed in the data. Validation loss also sees a general decrease from 0.08 to 0.03. However, both losses begin to increases from the 34th epoch onwards. The initial downward trend in model performance

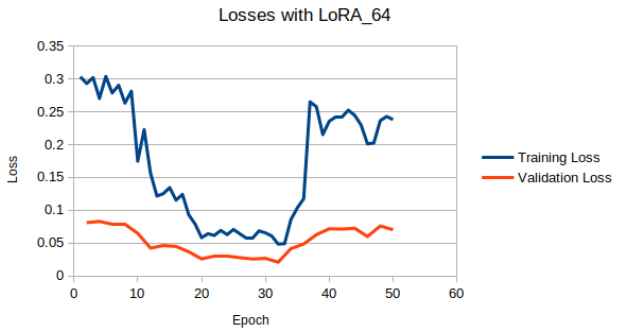

Figure 8: Performance of model with LoRA_64

lent some credibility to the optimisation procedure used but more

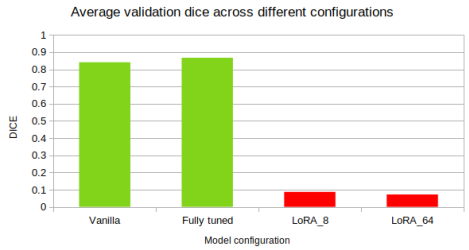

Figure 9: Plot of Dice scores for different configurations of SAM

scrutiny was required. Over-fitting is normally marked by a reducing training loss while validation loss increases. The lack of early stopping during optimization however means it is still a possibility.

Finally, using the Dice score to compare model performance was inconclusive. The Dice scores of the LoRA models completely deteriorated when compared to models without LoRA. This is evident by the distribution of values in Figure 9

5 CONCLUSIONS AND FUTURE WORK

SAM modified with LoRA adapters is able to be trained for better performance on tumour segmentation. However, performance comparisons between the vanilla SAM baseline and SAM with LoRA could not be reliably calculated

Further experimentation with different ranks of adapters and finer hyper-parameter tuning would allow more light to be shed onto the situation.

REFERENCES

[1] Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W. Remedios, Shunxing Bao, Bennett A. Landman, Lee E. Wheless, Lori A. Coburn, Keith T. Wilson, Yaohong Wang, Shilin Zhao, Agnes B. Fogo, Haichun Yang, Yucheng Tang, and Yuankai Huo. 2023. Segment Anything Model (SAM) for Digital Pathology: Assess Zero-shot Segmentation on Whole Slide Imaging. arXiv:2304.04155 [eess.IV]

[2] Lee R. Dice. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology* 26, 3 (1945), 297–302. http://www.jstor.org/stable/1932409

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV] https://arxiv.org/abs/2010.11929

[4] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2021. Masked Autoencoders Are Scalable Vision Learners. arXiv:2111.06377 [cs.CV] https://arxiv.org/abs/2111.06377

[5] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. arXiv:1902.00751 [cs.LG] https://arxiv.org/abs/1902.00751

[6] Bozhen Hu, Bin Gao, Cheng Tan, Tongle Wu, and Stan Z. Li. 2023. Segment Anything in Defect Detection. arXiv:2311.10245 [cs.CV]

[7] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL]

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. arXiv:2304.02643 [cs.CV]

[9] Dominic LaBella, Maruf Adewole, Michelle Alonso-Basanta, Talissa Altes, Syed Muhammad Anwar, Ujjwal Baid, Timothy Bergquist, Radhika Bhalerao, Sully Chen, Verena Chung, Gian-Marco Conte, Farouk Dako, James Eddy, Ivan Ezhov, Devon Godfrey, Fathi Hilal, Ariana Familiar, Keyvan Farahani, Juan Eugenio Iglesias, Zhifan Jiang, Elaine Johanson, Anahita Fathi Kazerooni, Collin Kent, John Kirkpatrick, Florian Kofler, Koen Van Leemput, Hongwei Bran Li, Xinyang Liu, Aria Mahtabfar, Shan McBurney-Lin, Ryan McLean, Zeke Meier, Ahmed W Moawad, John Mongan, Pierre Nedelec, Maxence Pajot, Marie Piraud, Arif Rashid, Zachary Reitman, Russell Takeshi Shinohara, Yury Velichko, Chunhao Wang, Pranav Warman, Walter Wiggins, Mariam Aboian, Jake Albrecht, Udunna Anazodo, Spyridon Bakas, Adam Flanders, Anastasia Janas, Goldey Khanna, Marius George Linguraru, Bjoern Menze, Ayman Nada, Andreas M Rauschecker, Jeff Rudie, Nourel Hoda Tahon, Javier Villanueva-Meyer, Benedikt Wiestler, and Evan Calabrese. 2023. The ASNR-MICCAI Brain Tumor Segmentation (BraTS) Challenge 2023: Intracranial Meningioma. arXiv:2305.07642 [cs.CV]

[10] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. 2023. Segment Anything in Medical Images. arXiv:2304.12306 [eess.IV]

[11] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. 2024. Segment anything in medical images. *Nature Communications* 15, 1 (Jan. 2024). https://doi.org/10.1038/s41467-024-44824-z

[12] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. arXiv:1612.00593 [cs.CV] https://arxiv.org/abs/1612.00593

[13] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. 2006. TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation. In *Computer Vision – ECCV 2006*, Aleš Leonardis, Horst Bischof, and Axel Pinz (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–15.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[15] Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. 2023. Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation. arXiv:2304.12620 [cs.CV]

[16] Kaidong Zhang and Dong Liu. 2023. Customized Segment Anything Model for Medical Image Segmentation. arXiv:2304.13785 [cs.CV] https://arxiv.org/abs/2304.13785