Computer Vision CS-GY 6643 - Final Project Report - Brain Tumor Semantic Segmentation

Daoyu Li, dl5312@nyu.edu,Ziming Song, zs2815@nyu.edu, Xuning Chang, xc1626@nyu.edu, Allison Lee, al6897@nyu.edu

Our repository: **link**

1 Introduction and background

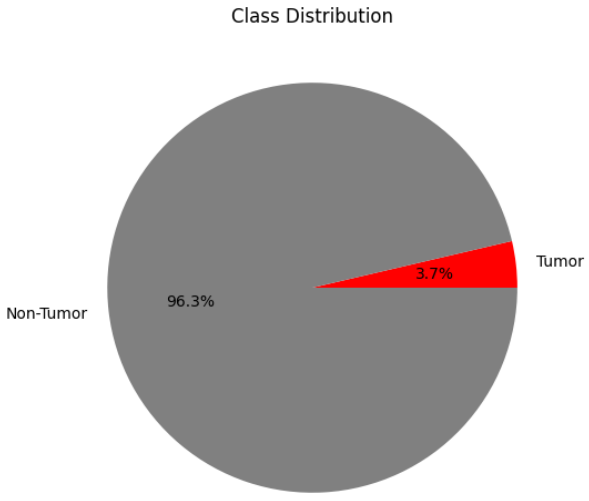
Brain tumor segmentation is a central task in medical image analysis, due to its importance for tumor diagnosis and later treatment. Traditional methods like U-Net (Walsh et al. 2022) have achieved a somewhat satisfactory result in medical image segmentation by utilizing an encoder-decoder architecture with skip connections to achieve high accuracy. However, the high labeling cost of U-net makes it difficult to generalize across diverse imaging conditions and tumor modalities. In many cases, it also struggles to output satisfactory masks.

In recent years, the Segment Anything Model (SAM) introduced by Meta AI has achieved great progress in the field of image segmentation. SAM is a foundation model for computer vision that demonstrates remarkable generalization across multiple domains with minimal task-specific training. SAM's powerful generalizing capability lessens the need for manual annotated datasets, which is expensive and required to train traditional segmentation models.(Kirillov et al. 2023) Building on this, SAM2 has been developed to enhance the accuracy of SAM. A recent study (Sengupta et al. 2024) has shown that SAM2 outperforms SAM in MRI-based computer vision tasks, offering superior accuracy and robustness even with limited training data. This finding promotes us to explore the application of SAM2 on brain tumor segmentation.

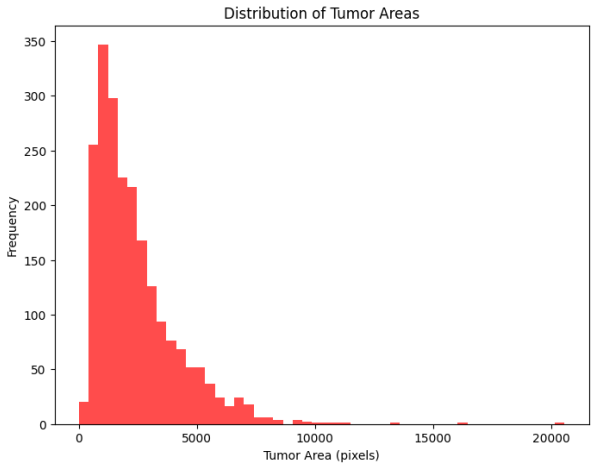
2 Datasets

We will conduct a few fine-tuning experiments on Brain Tumor Image Dataset: Semantic Segmentation (Seg 2023) from Kaggle. This dataset includes 2146 brain tumor images, separated into train, validation, and test set. An annotation is provided for each image, where it contains 2 classes: tumor (label 1) and not tumor (label 0).

The dataset is well-organized and the annotation format is in the standard COCO format. Not much pre-processing was necessary. MicroSAM requires a TIFF image as an input, so the conversion to such was the only pre-processing done. (see 3.3.1)



(a) Pixel distribution of dataset



(b) Tumor area on a single image

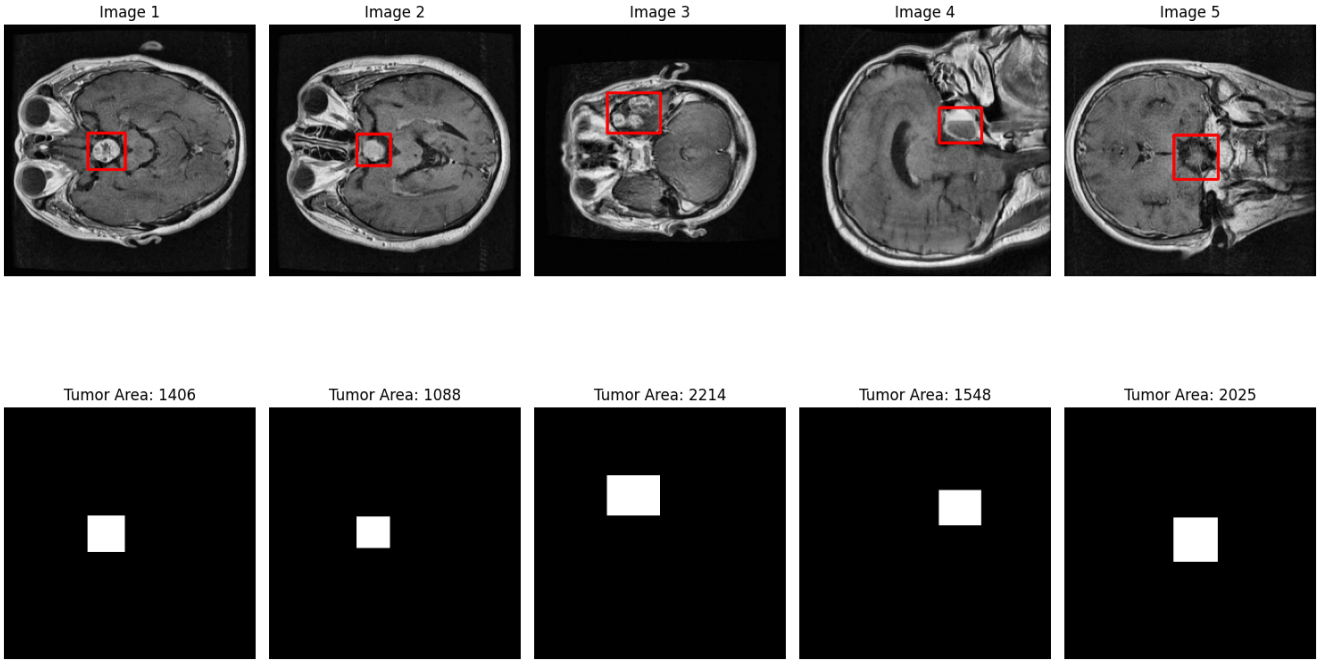


Figure 2: Example Images and Ground Truth Masks

3 Methods

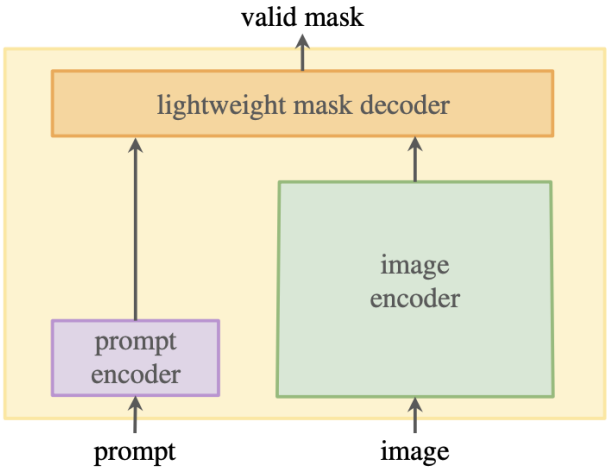


Figure 3: The architecture of SAM (Kirillov et al. 2023)

We attempted to fine-tune the Segment Anything Model 2 (SAM2) and microSAM to output a mask for each class (tumor or not tumor) for each image. SAM, SAM2's predecessor and microSAM's pre-fine-tuned counterpart, used a transformer architecture and trained on over 1 billion masks on 11 million images, which makes it possible to transfer zero-shot to new tasks with prompting. SAM2 improved upon SAM; it accelerated segmentation tasks and delivered more accurate results. MicroSAM (Archit et al. 2023) was fine-tuned on top of SAM with microscopy images. It is an efficient tool for segmentation and tracking in microscopy. By training specialized models for microscopy data, MicroSAM shows great potential in improving segmentation quality for a wide range of imaging conditions. Our goal is to build upon SAM (or SAM 2) and improve its performance in brain tumor segmentation by fine-tuning the parameters in its mask decoder (Figure 3) using our dataset.

Previous efforts to fine-tune SAM to medical images suggest fine-tuning the parameters from the encoder is not only computationally expensive, but often led to negative outcomes (Hu et al. 2023), so we only need to guide its decoding step where the model transform the encoded input image into the output mask by fine-tuning the decoder.

To train the decoder, we use the dice loss function with the Adam optimizer with standard hyper-parameters. The dice loss function measures the mask's intersection area, where let Y and Y' be the ground truth mask and the model output mask respectively,

$$Dice(Y, Y') = \frac{2(Y \cap Y')}{Y + Y'}$$

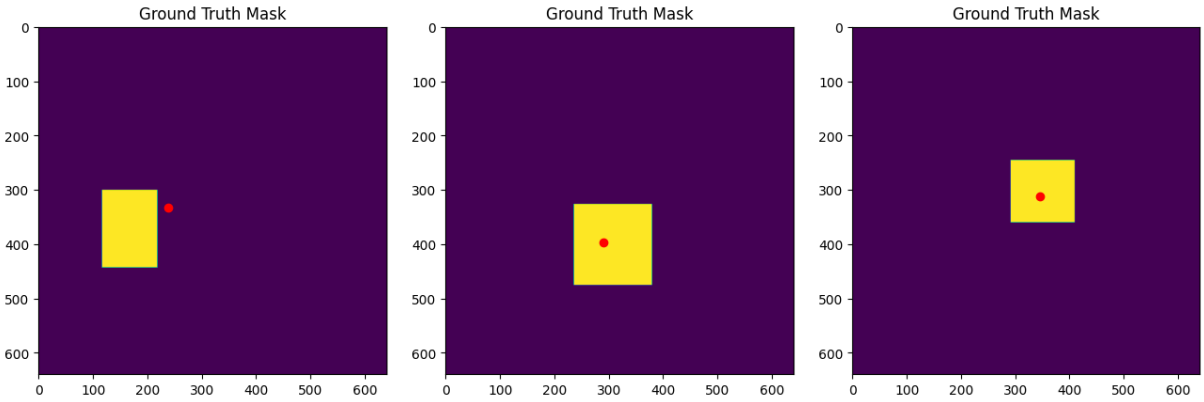
We will follow the approach taken by Hu et al., where we will freeze the weights of the encoder and disable the prompting encoder. (Hu et al. 2023) Their approach fits well in the context of segmenting brain tumor images.

3.1 SAM2

3.1.1 Prompt Generation

Before we can train the model, SAM2 requires either a bounding box or a point prompt to generate a mask. In other words, during our training, we also need to provide prompts to the model for it to infer a mask, and thus fine-tune the model using such masks.

We generate these prompts by using the centroids of the ClipSeg mask. As mentioned in the previous sections, this is the model with the best results so far. We reproduced the model using the provided notebook (with a few adjustments to debug). Then, we calculate the centroid of the masks produced by this model to give us one point to prompt SAM2. In most instances, it was able to give us a point located inside (or at least very close to) the ground truth mask.



3.1.2 Data Preprocessing

We first generated ground truth mask from the COCO annotation file, and prompt points from ClipSeg mask respectively for all images. Then, we loaded all of them into a PyTorch dataset object for training.

3.1.3 Training and Fine-tuning

During training, only the mask decoder parameters were updated, while the encoder weights remained frozen to reduce computational cost. Initially, we used the Adam optimizer with a learning rate of $1e^{-5}$, but it did not perform well; specifically, the loss stopped decreasing after the first epoch. To address this issue, we switched to the AdamW optimizer with the same learning rate, which led to better convergence and improved performance. The Dice loss function was used for optimization.

We experimented with three SAM2 model sizes: SAM2 Small, SAM2 Base Plus, and SAM2 Large. All models were trained for 5 epochs, with gradient updates applied every 4 steps. Among these, SAM2 Large achieved the best performance with an IoU of 30.17%, surpassing the zero-shot baseline of 28.50%.

3.2 AutoSAM

AutoSAM (Shaharabany et al. 2023) builds upon SAM and introduces a lightweight segmentation solution by adding an encoder and without further fine-tuning SAM. This new encoder is trained via gradients provided by a frozen SAM.

In our work, we adapt the zero-shot AutoSAM basic model to our dataset. For each image, we use ground truth to generate a single point as a prompt, making sure the correctness of prompt. Then we use AutoSAM `sam_vit_b` checkpoint to predict the segmentation mask, and evaluated the results accordingly.

Due to its limited mean IoU performance in the zero-shot setting and the challenges in reproducing its fine-tuning code, we ultimately chose not to perform fine-tuning on AutoSAM.

3.3 MicroSAM

MicroSAM (Archit et al. 2023) as Segment Anything for Microscopy implements automatic and interactive annotation for microscopy data. It is built on top of Segment Anything by Meta AI and specializes in microscopy and other biomedical imaging data.

3.3.1 Dataset Preprocessing

We first generated ground truth masks from the provided COCO annotation files and store it into runtime memory. Then, to make them compatible with microSAM data loaders, we convert the current PNG image files and store the mask image to TIFF (.tif) format. These preprocessing steps ensure the dataset would be usable during training.

3.3.2 Dataloader Configuration

The preprocessed dataset is split into three parts: a training set, a validation set, and a test set, according to the folder split given by the dataset. This splitting ensures that the model can be properly trained, validated without overfitting, and can be generalized to unseen samples. The dataloader is configured to load the training set and validation set for the fine-tuning process.

3.3.3 Training and Fine-Tuning

We initially fine-tuned the microSAM ViT-B model using the dataloaders and achieved a mean IoU score of 0.5019 on the test set. During analysis, we observed that the ViT-B model generated multiple masks, which contributed to the lower mIoU score. Then we implement a function to find the most likely class from the generated masks and let the model generate a single-class mask. This adjustment slightly improved the mIoU score to 0.5042.

From this result, we realize that the model is limited by the overall ability to generate masks. To improve segmentation performance, we fine-tuned a larger model, microSAM ViT-H, which finally performs better and has the mIoU score of 0.5954. This demonstrates the benefit of using larger models to improve the accuracy of brain tumor segmentation.

We also tried to fit the predicted masks with the smallest encircling rectangle, to align with the ground truth mask. However, after applying this modification, the mean IoU is decreased to 0.5544. Therefore, we decided to keep using the generated masks without modification on them.

4 Baseline Methods

In the dataset's hosting websites (Kaggle and Roboflow), we see many previous attempts to this problem by fine-tuning other more well-known models, including U-Net and R-CNN. (*Brain Tumor Image DataSet : Semantic Segmentation — kaggle.com* n.d.)

The best model (shown in Roboflow) claimed a dubious mIoU score of 66% after fine-tuning. However, this number was not backed up by any sources, nor did the author provide their source code. Thus, this result should be discarded. The next best model has an mIoU score of 56% by fine-tuning CLIPSeg.(*CLIPSeg — kaggle.com* n.d.) CLIPSeg is another transformer-based segmentation model in 2021, although its influence was eclipsed by the newer SAM and SAM2 models.(Lüddecke et al. 2022)

For this project, we have selected SAM2 as our baseline method because the newer transformer architecture promises a better segmentation result and could be more generalizable than traditional architectures like U-Net or R-CNN. With an unprecedented amount of foundational model training data, we could also expect SAM2 to perform much better than other transformer-based models like CLIPSeg.

Although there is no existing research specifically on SAM2 for brain tumor segmentation, SAM has demonstrated robust generalization capabilities and effective segmentation performance in brain tumor semantic segmentation (Zhang et al. 2024) and other downstream tasks in the medical industry. Therefore, SAM2 is chosen as the baseline to explore its effectiveness in the context of semantic segmentation of brain tumors, with the hope of improving the precision of segmentation in terms of mIoU.

5 Evaluation

Following the industry standard in evaluating image segmentation model, we will use the mean Intersection over Union (mIoU) as our evaluation metric. The mIoU score has long been seen as the gold standard in evaluating semantic segmentation models. To calculate this, for each class on an image, we calculate the quotient between intersectional area of the ground truth mask and the model output and the union, or the combination, of those areas. (Bernhard et al. 2024) In other words, let Y =ground truth mask, Y' = model mask,

$$IOU(Y, Y') = \frac{Y \cap Y'}{Y \cup Y'}$$

Essentially, the more similar it is between the ground truth and the model output, the higher the score would be. Then, the mIoU score would be the average IoU score of each class mask in each image.

In the evaluation notebook, we displayed the results of zero-shot SAM2 (baseline) and finetuned microSAM. The results for fine-tuned SAM 2 is not currently ready for display.

All models share the same evaluation flow by calling the `evaluate` function and provide an iterator for the dataset, predictor function, and optionally the transformation function (for tasks like scaling). This evaluation function automatically displays the visualization of the image, ground truth mask, and the predicted mask for the first three samples. In the end, it would also return the mean IoU score for the test dataset.

6 Results

Table 1 summarizes segmentation performance of methods we experimented in terms of mean IoU. The baseline SAM2 zero-shot method achieves a Mean IoU of 28.50%, while fine-tuned SAM 2 and zero-shot AutoSAM slightly improves upon this to 30.17% and 31.62% respectively. Fine-tuned ClipSeg significantly boosts performance to 56.10%, and the current best performance is achieved by fine-tuning microSAM, yielding a Mean IoU of 59.54%.

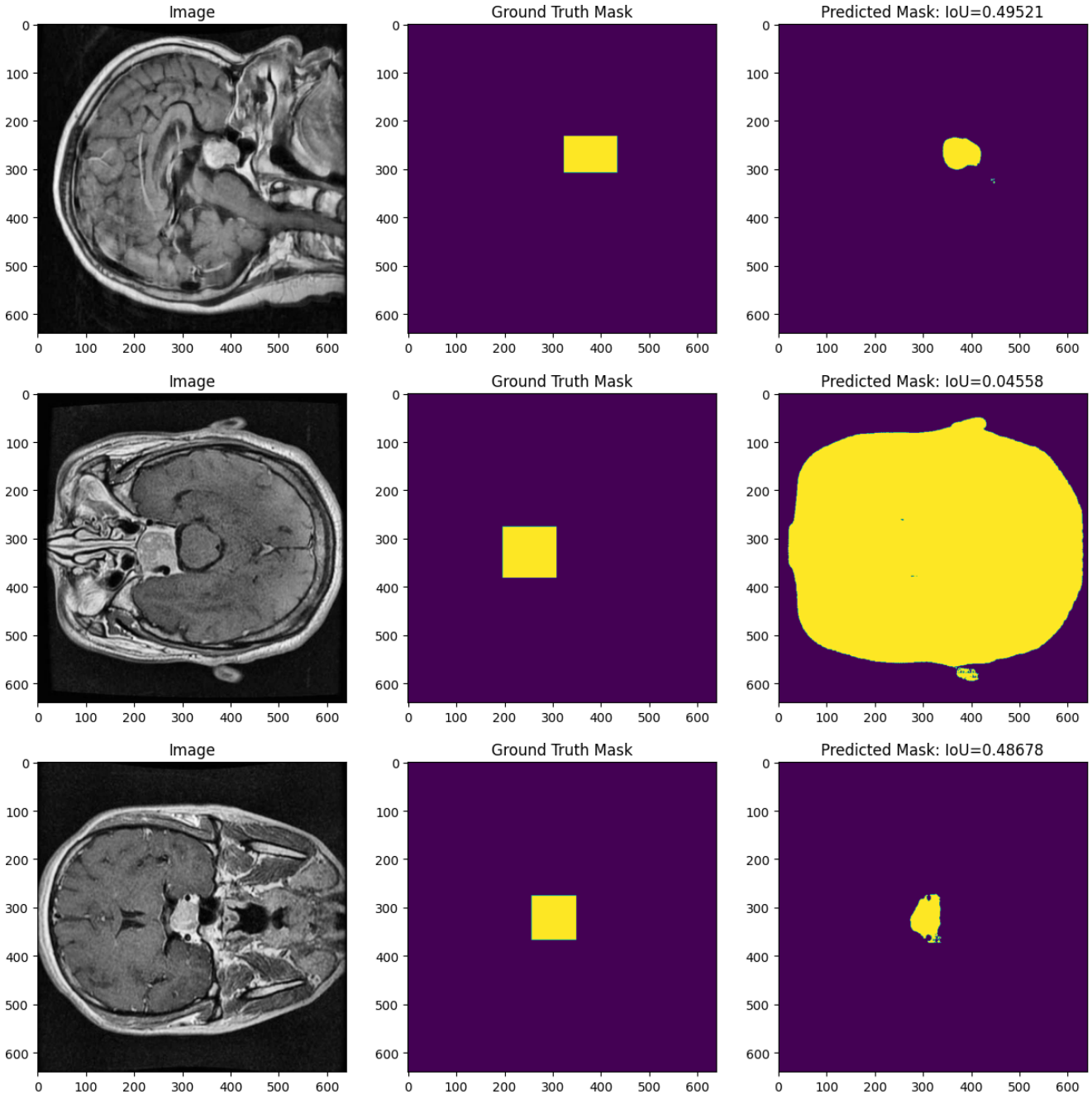
MicroSAM performed the best on this task because it was fine-tuned with microscopy data, which aligned with our dataset. Brain tumor images resemble MRI-like snapshots and share structural similarities with microscopy imaging data. By training on such a domain-specific dataset, MicroSAM can capture nuanced patterns that vanilla SAM2 might overlook.

Method	Mean IoU
SAM2 zero-shot (baseline)	**28.50%**
Finetuned SAM2	**30.17%**
AutoSAM zero-shot	**31.62%**
Finetuned ClipSeg (Kaggle's best)	**56.10%**
Finetuning microSAM (our best)	**59.54%**

Table 1: Results

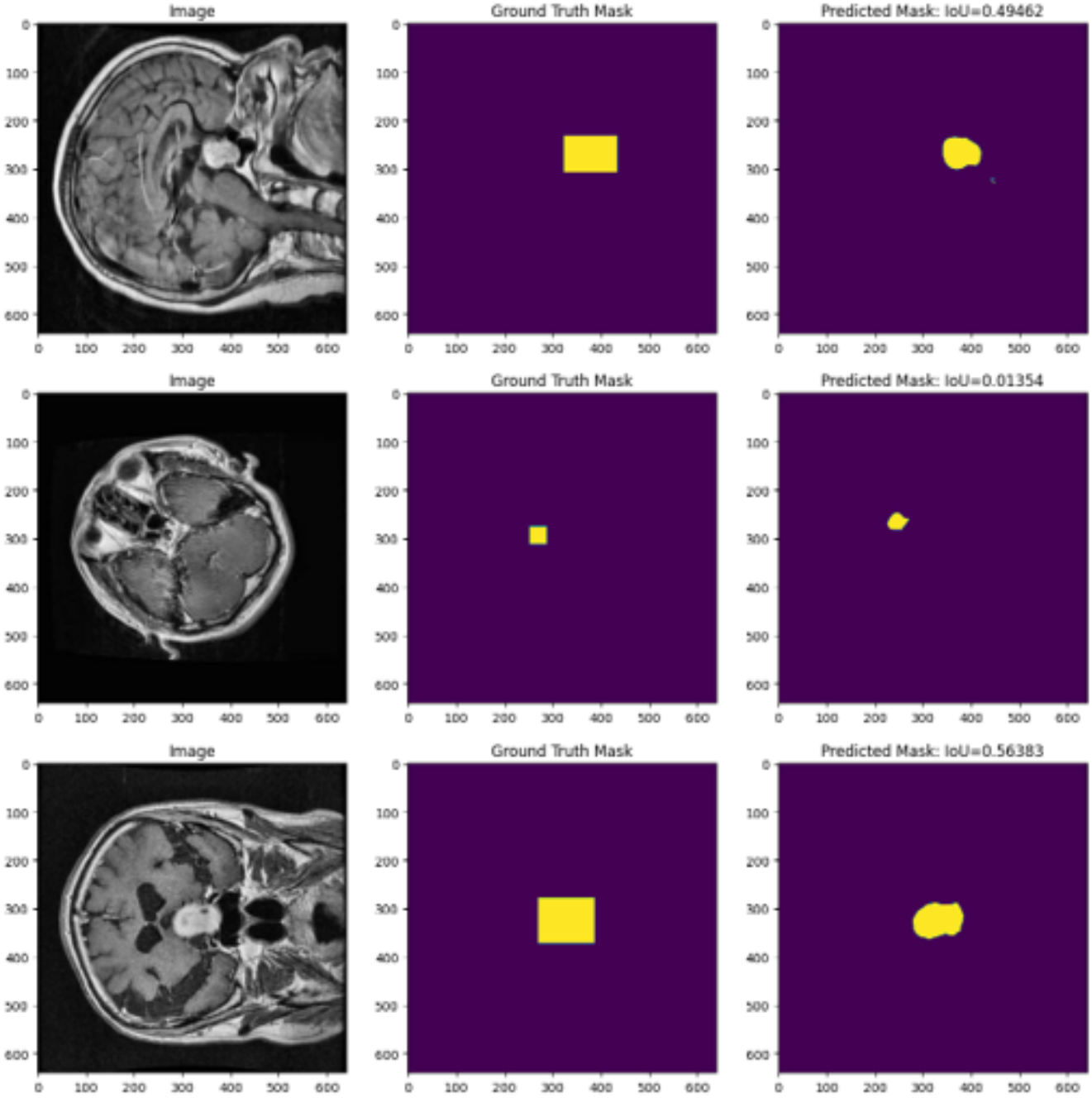
6.1 Baseline

Below are the masks predicted by zero-shot SAM2 from 3 sample images in the test dataset.



6.2 Fine-tuned SAM2

Below are the masks predicted by fine-tuned SAM2 for 3 sample images in the test dataset. The best-performing model is trained on SAM2 Large, achieving a mean IoU of 30.17%, which surpasses the baseline SAM2's zero-shot performance of 28.5%. Initially, we experimented with SAM2 Base Plus, however, it underperformed the basleine with a mean IoU of 24.5%. While SAM2 Large showed improvement, some predicted masks still failed to align precisely with the ground truth.



6.3 Fine-tuned microSAM

Compared to the baseline, the fine-tuned microSAM model has a significant improvement in segmentation performance. Its mean IoU score increased to 59.54%, which outperforms the baseline SAM2's zero-shot performance of 28.50% and the fine-tuned ClipSeg model's 56.10%. Beyond the numerical metrics, the predicted masks from the fine-tuned microSAM have a higher precision in capturing tumor regions, which were more similar to the ground truth masks.

However, from the last two test images shown below, the generated mask is either much larger than the ground truth or completely missed the desired area. These two conditions could be the culprit of mean IoU score not being higher.

We observed that the larger MicroSAM ViT-H model performed better than its smaller ViT-B variant. This outcome matched with the general observation that larger model sizes can more effectively capture the complex features required for accurate brain tumor segmentation.

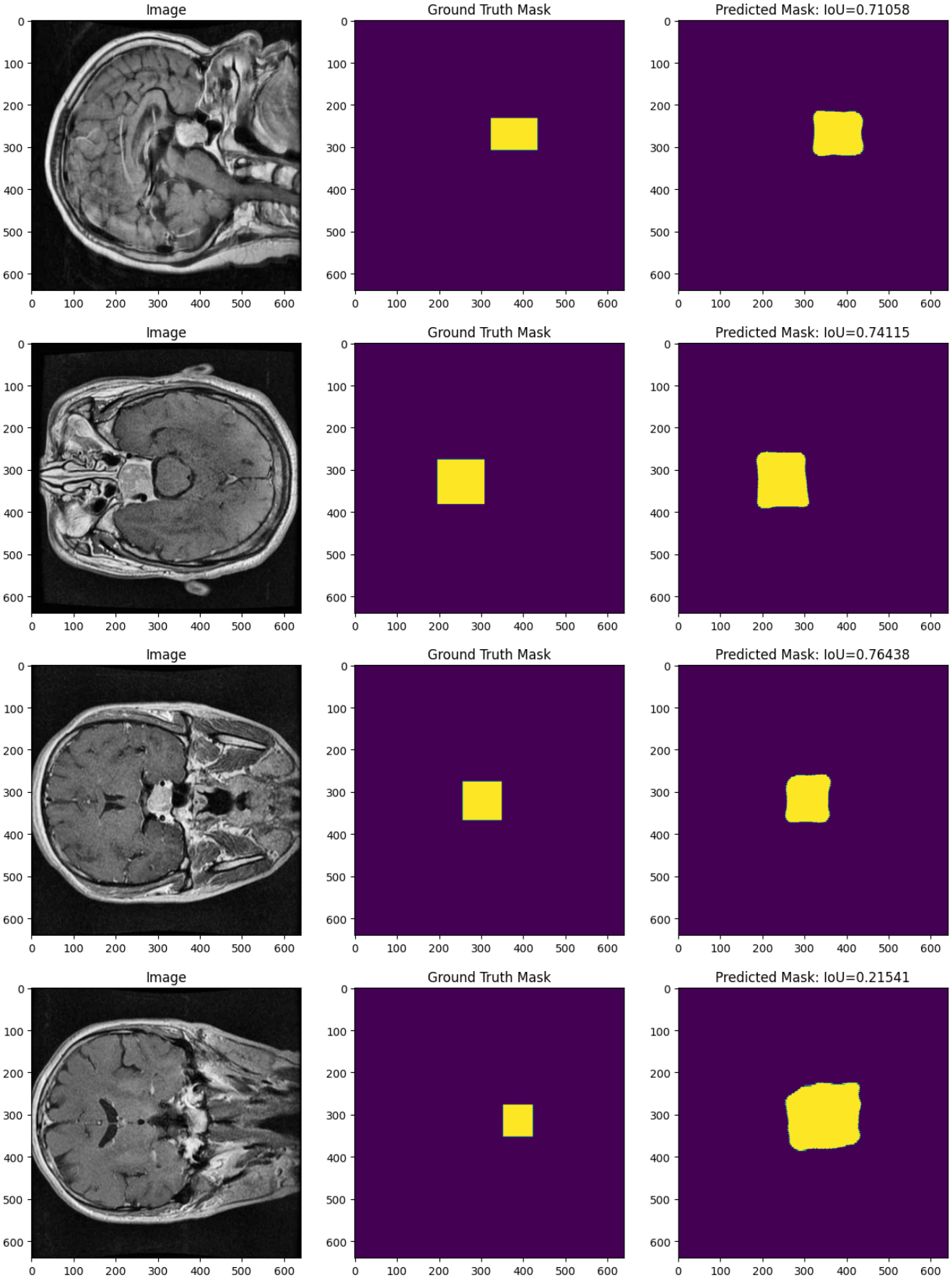
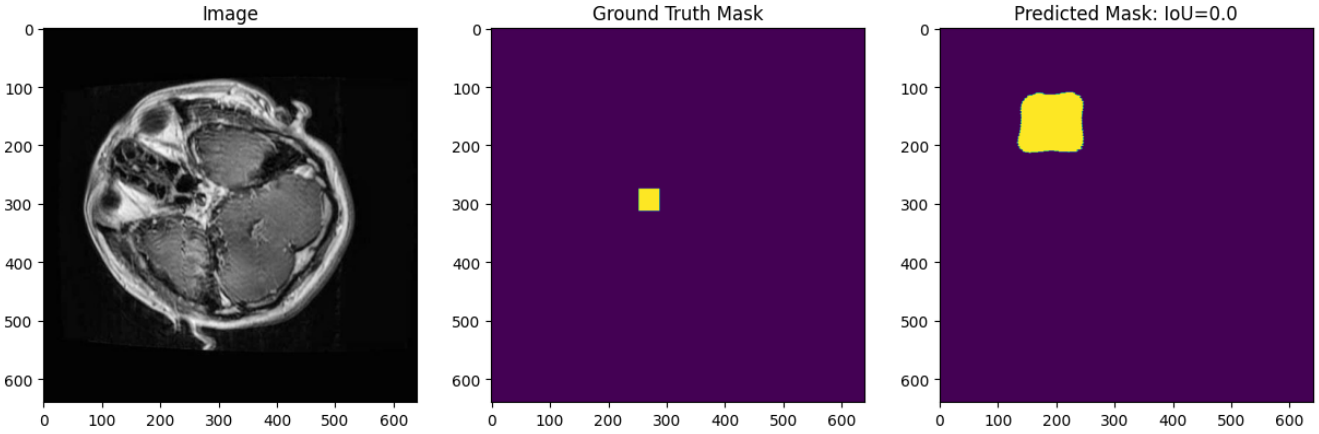


Image
Ground Truth Mask
Predicted Mask: IoU=0.71058
Image
Ground Truth Mask
Predicted Mask: IoU=0.74115
Image
Ground Truth Mask
Predicted Mask: IoU=0.76438
Image
Ground Truth Mask
Predicted Mask: IoU=0.21541



7 Author contributions

	Daoyo Li	Ziming Song	Xuning Chang	Allison Lee
Prompt Generation				■
AutoSAM Fine-tuning (failed)		■		
microSAM Fine-tuning	■	■		
SAM 2 Fine-tuning			■	
Evaluation	■	■	■	■
Report Writing	■	■	■	■

References

Archit, A. et al. (2023). "Segment Anything for Microscopy". In: *bioRxiv*. DOI: 10.1101/2023.08.21.554208. eprint: https://www.biorxiv.org/content/early/2023/08/22/2023.08.21.554208.full.pdf. URL: https://www.biorxiv.org/content/early/2023/08/22/2023.08.21.554208.

Bernhard, M. et al. (2024). "What's Outside the Intersection? Fine-Grained Error Analysis for Semantic Segmentation Beyond IoU". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 968–977.

Brain Tumor Image DataSet : Semantic Segmentation — kaggle.com (n.d.). https://www.kaggle.com/datasets/pkdarabi/brain-tumor-image-dataset-semantic-segmentation/code. [Accessed 12-10-2024].

CLIPSeg — kaggle.com (n.d.). https://www.kaggle.com/code/phucvr/clipseg. [Accessed 12-10-2024].

Hu, X., X. Xu, and Y. Shi (2023). *How to Efficiently Adapt Large Segmentation Model(SAM) to Medical Images*. arXiv: 2306.13731 [cs.CV]. URL: https://arxiv.org/abs/2306.13731.

Kirillov, A. et al. (2023). *Segment Anything*. arXiv: 2304.02643 [cs.CV]. URL: https://arxiv.org/abs/2304.02643.

Lüddecke, T. and A. S. Ecker (2022). *Image Segmentation Using Text and Image Prompts*. arXiv: 2112.10003 [cs.CV]. URL: https://arxiv.org/abs/2112.10003.

Seg (2023). *TumorSeg Dataset*. https://universe.roboflow.com/seg-aokuq/tumorseg. Open Source Dataset. visited on 2024-10-10. URL: https://universe.roboflow.com/seg-aokuq/tumorseg.

Sengupta, S., S. Chakrabarty, and R. Soni (2024). *Is SAM 2 Better than SAM in Medical Image Segmentation?* arXiv: 2408.04212 [eess.IV]. URL: https://arxiv.org/abs/2408.04212.

Shaharabany, T. et al. (2023). *AutoSAM: Adapting SAM to Medical Images by Overloading the Prompt Encoder*. arXiv: 2306.06370 [cs.CV]. URL: https://arxiv.org/abs/2306.06370.

Walsh, J. et al. (2022). "Using U-Net network for efficient brain tumor segmentation in MRI images". In: *Healthcare Analytics* 2, p. 100098. ISSN: 2772-4425. DOI: https://doi.org/10.1016/j.health.2022.100098. URL: https://www.sciencedirect.com/science/article/pii/S2772442522000429.

Zhang, P. and Y. Wang (2024). *Segment Anything Model for Brain Tumor Segmentation*. arXiv: 2309.08434 [eess.IV]. URL: https://arxiv.org/abs/2309.08434.