UNIVERSITY OF CAPE TOWN

DEPARTMENT OF COMPUTER SCIENCE

# CS Honours Project Final Paper 2024

Title: **U-SAM for Brain MRI: A Hybrid Approach for Intracranial Meningioma Segmentation with SAM and U-Net**

Author: Cassandra Wallace

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** | |

# U-SAM for Brain MRI: A Hybrid Approach for Intracranial Meningioma Segmentation with SAM and U-Net

Cassandra Wallace
University of Cape Town
Cape Town, South Africa
wllcas004@myuct.ac.za

## ABSTRACT

Medical Image Segmentation (MIS), particularly for brain tumours such as intracranial meningiomas, is critical for accurate diagnosis and treatment planning. The Segment Anything Model (SAM), while effective in natural image segmentation, faces challenges when applied to medical imaging due to the need to convert 3D grayscale MRIs into 2D RGB slices, leading to information loss. This paper explores fine-tuning SAM with a slice-by-slice approach and integrating it with a U-Net architecture to improve segmentation accuracy. We present a framework, U-SAM, that combines the strengths of both models. Our results demonstrate that while the fine-tuned SAM shows notable improvements, the integrated U-SAM framework offers potential advancements in segmentation precision, particularly when combining predictions from both models.

## KEYWORDS

Computer Vision, Deep Learning, Medical Image Segmentation, Tumour Segmentation, Segment Anything Model

## 1 INTRODUCTION

Medical Image Segmentation (MIS) plays a pivotal role in modern healthcare by enabling the accurate delineation of anatomical structures within medical images. This level of precision is essential for effective diagnosis, treatment planning, and ongoing disease monitoring [30]. Traditional segmentation methods, often rely on manual processes that are time-consuming and prone to human error [19].

Deep Learning models, such as the Segment Anything Model (SAM), offer a promising alternative, demonstrating impressive segmentation capabilities in various domains [23]. However, SAM was designed primarily for 2D natural images, which poses significant challenges when applied to 3D medical imaging tasks, particularly those involving brain Magnetic Resonance Imaging (MRI) scans. The primary challenges include domain-specific differences such as lower contrast, higher noise levels, and the intricate nature of deep anatomical structures [9, 16, 30].

Brain MRIs are crucial for the visualisation of brain structures, which is essential for diagnosing and treating intracranial meningiomas, the most prevalent type of brain tumour in adults [22, 24]. While meningiomas are frequently benign, their potential for aggressive behaviour underscores the need for precise segmentation to support effective treatment planning [18, 32].

The goal of this research is to adapt SAM for the task of 3D brain MRI segmentation by fine-tuning it on the 2023 BraTS Intracranial Meningioma Challenge dataset [24]. This adaptation involves addressing challenges to effectively leveraging SAM's capabilities to improve its performance in medical imaging tasks. Additionally, we explore whether integrating SAM with a U-Net architecture—a model renowned for its success in biomedical image segmentation [39]—can enhance its performance in segmenting intracranial meningiomas, relative to baseline SAM.

The design of our approach involves a multi-faceted strategy. Initially, we fine-tune SAM on the dataset to adapt its parameters and improve its performance in 3D medical imaging. We then integrate SAM with U-Net to leverage the strengths of both models. Each step of the design process is carefully justified based on the unique challenges posed by 3D MRI data, the capabilities of SAM and U-Net, and insights gained from existing efforts. The design decision to integrate these models is driven by their complementary strengths—SAM's broad applicability and U-Net's proven success in biomedical segmentation. The ingenuity in our approach lies in effectively combining these models to tackle the specific challenges of intracranial meningioma segmentation.

The remainder of this paper is structured as follows: Section 2 and Section 3 provide a comprehensive overview of the background and related work of the field, including existing attempts to adapt SAM to MIS. Section 4 details our chosen dataset and the proposed modifications for integrating SAM with U-Net [39]. Section 5 presents and discusses our experimental results, comparing our modified model's performance with the baseline SAM, followed by a summary in Section 6. Finally, Section Section 7 concludes the paper with a discussion of our findings and potential future work in this area. Detailed visualisations are included in the Appendices, which provide further insight into the performance and behaviour of the models discussed.

## 2 BACKGROUND

This section provides an overview of the evolving landscape of image segmentation, from traditional Machine Learning methods to advanced Deep Learning techniques that have transformed MIS, such as U-Net and Vision Transformers. By highlighting the shift from hand-crafted features to automated feature learning, we contextualise the advancements shaping current models, such as SAM, and discuss their relevance to MIS.

### 2.1 Machine Learning

Machine Learning has been foundational in medical imaging, but is limited by its reliance on hand-crafted features. Classical approaches, such as simple thresholding [20] or more advanced clustering-based algorithms [14], often struggle with the complexity of medical data, particularly in effectively capturing the inherent intricate patterns.

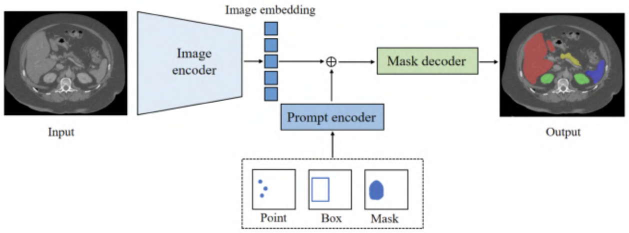


**Figure 1: The SAM Architecture, with Image Encoder, Prompt Encoder, and Mask Decoder [23].**

## 2.2 Deep Learning

Deep Learning architectures, often consisting of encoder-decoder structures with millions of parameters, have revolutionised image segmentation by automatically extracting meaningful features to learn and understand the complex relationships within image data.

In particular, Convolutional Neural Networks (CNNs) have become significantly influential [5]. They leverage deeply layered architectures to learn hierarchical representations of images, enabling the identification of complex patterns that are critical in medical imaging. Their ability to accurately capture spatial relationships has made them the backbone of many medical imaging systems [5]. However, CNNs are inherently limited by their local receptive fields, which restrict their capacity to capture global image context.

One of the most impactful architectures in this domain is U-Net [39], known for its symmetric design. U-Net consists of a contracting path to capture context and a symmetric expanding path that enables precise localisation. This architecture has been extensively applied to various medical imaging tasks, including brain tumour segmentation, because of its robustness and accuracy in handling the intricate details of medical images.

## 2.3 Visual Transformers

Transformers, originally developed for Natural Language Processing, have emerged as a powerful alternative to CNNs in image processing. Vision Transformers (ViTs) [10], utilising self-attention mechanisms, can capture global image dependencies, making them highly effective for tasks requiring broader context. This advantage, not provided with CNNs as they rely on local convolutional operations to process images [5], has translated well to MIS, where capturing entire anatomical structures is critical [10].

The Segment Anything Model (SAM), introduced by Kirillov et al. in 2023 [23], represents a significant advancement in image segmentation. As seen in Figure 1, it builds upon the ViT architecture, consisting of three components: an image encoder, a prompt encoder, and a mask decoder.

The image encoder processes the input image into a high-dimensional image embedding space using a Vision Transformer (ViT) pre-trained with the Masked Auto-Encoder (MAE) training scheme [15]. The ViT efficiently captures and extracts essential features and relationships within the image. Meta AI provides three pre-trained SAM models corresponding to the three ViT sizes (ViT-Base, ViT-Large, and ViT-Huge) [10]. However, their research suggests that using larger ViT models as the backbone of SAM's image encoder offers only marginal improvements in accuracy, while significantly increasing computational demands [23].

SAM's prompt encoder translates user prompts into internal representations that the model can understand. These prompts can be in the form of points (positive or negative, to indicate the foreground and background of the target respectively), bounding boxes (to surround the spatial region of the target object), or even textual descriptions. SAM also offers a special 'Everything' mode, where the model segments all potential objects in the entire image [23].

The mask decoder iteratively combines the image embeddings from the image encoder and the prompt embeddings from the prompt encoder to generate segmentation masks. This process reflects the input image's visual features and the user's specified target through the prompt. The decoder itself is designed as a modified Transformer decoder block, utilising techniques such as prompt self-attention, which allows the prompt embeddings to interact and refine their representations, and cross-attention in two directions, which facilitates information exchange between the image embedding and the prompt embeddings [23].

This design allows SAM to perform zero-shot learning across many images without requiring task-specific training, generating accurate masks without exposing explicit object-level data. While SAM's versatility is impressive, particularly between natural domains [21, 40], its application to medical images, especially 3D medical imaging, presents unique challenges, even with its zero-shot learning. Medical images often involve lower contrast, higher noise levels, and 3D anatomical structures, which complicates the direct application of SAM's original design [9, 16].

# 3 RELATED WORK

Research on adapting SAM for medical imaging has largely focused on addressing its limitations in handling 3D data and fine-tuning prompt mechanisms for more precise segmentation. This section categorises key contributions, including modifications to handle 3D data, advancements in prompting methods, fine-tuning approaches, and existing integration efforts.

## 3.1 MedSAM

MedSAM [30] was one of the earliest contributions to the adaptation of SAM for medical imaging. By tailoring SAM for many medical domains, MedSAM aimed to create a universal model capable of handling different medical imaging tasks. A key innovation was the use of bounding box prompts and the handling of 3D data through image slicing. This method involved slicing 3D volumes into 2D sections and adopting a full fine-tuning approach. The model demonstrated improved performance across various cancer imaging modalities, particularly in accurately capturing the intricate details of tumour boundaries and accommodating the varying nature of medical datasets. MedSAM's success laid the foundation for subsequent research, influencing many later adaptations.

## 3.2 3D to 2D Approaches

Many approaches to adapting SAM for 3D MIS involve tackling the inherent complexity of 3D data. A common strategy has been to convert 3D volumes into 2D slices—a technique known as slice-by-slice segmentation, as seen in MedSAM [30]. While this approach simplifies the application of SAM to 3D data, it can lead to the loss of critical contextual information inherent in 3D structures [42].

To address this, several studies have investigated advanced methods that adapt SAM's input to process complete 3D volumes. For example, certain adaptations override SAM's 2D requirement by incorporating a 3D encoder and slicer, such as SAM3D [3], MA-SAM [6], and 3DSAM-Adapter [13]. More minimally, SlideSAM [36] addressed the limitations of slice-by-slice segmentation by utilising a windowing approach to capture contextual information.

Although this approach allows SAM to process 3D data more effectively, it requires substantial computational resources and remains a complex area of research [6]. While our research primarily focuses on tuning SAM's ability to handle 2D slices, understanding these 3D approaches is important for contextualising the broader challenges in the field.

## 3.3 Prompting Method Approaches

In adapting SAM for medical imaging, the choice of prompting method is a critical factor. Most approaches have employed manual prompts, such as bounding boxes or points, to guide the segmentation process. While effective, these methods require human intervention, which can be a bottleneck in clinical settings where medical expertise is needed to provide accurate prompts [42].

Ma et al. [30] highlighted the limitations of point-based prompting, noting that it often leads to ambiguity and may require multiple iterations to achieve accurate segmentation. On the other hand, bounding box prompts are argued to be more effective in specifying the target object with minimal human intervention. However, even bounding boxes can struggle with complex structures, such as vessel-like branches, where enclosing the entire structure might inadvertently include irrelevant regions due to the inherent shape of the target [42].

In contrast, SAM's fully automated *"Everything"* prompting mode, which segments all potential objects within an image, eliminates human input but may lack the precision required for specific medical tasks. This mode can result in segmentations that are too broad, missing the focused regions of interest that medical professionals typically require.

Other prompting methods, such as text prompts, have been explored in studies like MedSAM [30], which developed a version incorporating text-based cues. However, text prompting is less relevant for clinical use, where accurate spatial localisation is more critical than abstract descriptive cues. Our research focuses on using bounding box prompts to optimise SAM's performance in brain MRI segmentation.

## 3.4 Fine-Tuning Approaches

One of the primary methods for adapting SAM to medical imaging is full fine-tuning, where all components of SAM—image encoder, prompt encoder, and mask decoder—are retrained on a medical dataset. This approach leverages the pre-trained weights of SAM as a foundation, allowing the model to adapt more effectively to the nuances of medical images.

For instance, SkinSAM [17], fully fine-tuned SAM for skin cancer segmentation. Their model achieved a notable performance improvement, with accuracy scores increasing from 81% to 89%, particularly excelling in segmenting vascular lesions. Similarly, Polyp-SAM [27], fully fine-tuned SAM for polyp segmentation, yielding consistent accuracy scores above 88% across multiple datasets, showcasing the effectiveness of this approach in diverse medical contexts. These examples demonstrate that full fine-tuning can achieve high accuracy; however, this approach often requires significant computational resources and extended training times, which can be a limitation in practical applications.

To address the challenges of computational cost and training time inherent in full fine-tuning, some researchers have explored Parameter-Efficient Fine-Tuning (PEFT) methods. PEFT approaches selectively fine-tune only a subset of the parameters of SAM, thereby balancing the benefits of pre-trained knowledge with the need to adapt to medical images [11]. MedSAM [30] fine-tuned SAM's image encoder and mask decoder while keeping the prompt encoder frozen. This method focused on adapting SAM to medical data without altering its core ability to understand prompts, resulting in a model that maintained strong generalisability while achieving high accuracy. This popular PEFT-SAM strategy not only reduces computational burden but also preserves the inherent versatility of SAM, allowing it to be more easily applied across various medical domains [11, 30, 33].

## 3.5 Framework Modification Approaches

Another significant direction in adapting SAM to the medical domain involves modifying its framework or integrating it with other established architectures [42].

For instance, nnSAM, proposed by Li et al. [28], integrates SAM with nnU-Net [19], a modern and highly flexible segmentation framework. Integrating SAM as a plug-and-play module within the nnU-Net architecture, nnSAM delivers superior segmentation accuracy over standalone SAM or nnU-Net. This integration highlights how SAM's strengths in general segmentation can be complemented by nnU-Net's specialised handling of medical images.

Similarly, ClipSAM [26] incorporated SAM with CLIP (Contrastive Language-Image Pre-training), a model developed by OpenAI that aligns images with text descriptions [37], enabling powerful zero-shot image recognition and multi-modal capabilities. This integration enhances SAM's versatility in handling various text-prompted segmentation tasks, positioning it as a robust tool for multi-modal medical image analysis [26].

Another innovative approach is SAMUS, developed by Lin et al. [29]. It introduces a parallel CNN branch that injects local features directly into the image encoder of SAM. This enhancement enriches the feature representation, making SAMUS particularly well-suited for handling smaller input sizes, which are common in medical imaging. SAMUS has been shown to outperform both MedSAM [30] and other adaptations across various medical imaging modalities, while significantly reducing computational costs [29].

However, while effective, these framework modifications often introduce additional complexity, which may affect their ease of deployment in real-world healthcare settings [42].

Despite these advancements, the adaptation of SAM for MIS is still evolving. Although many new studies have emerged during the timeline of our research, several challenges persist in adapting SAM for 3D brain MRI segmentation. Most notably, the transition from 3D to 2D slicing can result in the loss of critical contextual information, affecting the accuracy of the segmentation. Furthermore, while fine-tuning SAM's mask decoder has shown promise, it is still resource-intensive and may not always yield significant improvements. In addition, many of the current adaptations require extensive computational resources, which limits their accessibility and scalability.

Our study aims to contribute to this growing research by investigating an approach to adapt SAM for intracranial meningioma segmentation in brain MRI scans. We seek to address the existing gaps by comparing the performance of our modified models to the original SAM model, intending to achieve marginal improvements in accuracy and efficiency.

## 4 DESIGN AND IMPLEMENTATION

This section outlines the experimental setup, detailing the dataset, pre-processing methods, and modified model architectures. All design choices and methodologies are presented to ensure reproducibility and robustness in evaluating our proposed U-SAM-Combo and U1-SAM2 architectures for intracranial meningioma segmentation. The rationale behind our approaches is informed by the prior work discussed in Section 2 and Section 3. Many of these design choices were influenced by Kurtlab's BraTS 2023 submission [38], which provided valuable guidance in refining our strategies and navigating the associated challenges.

### 4.1 Data Collection

While many high-quality, annotated MRI datasets exist [35, 41], our research utilised the BraTS 2023 Intracranial Meningioma dataset, which includes MRI scans across four different types of modalities: T1 Weighted Native (t1n), T1-Contrast Enhanced (t1c), T2-Weighted Fluid Attenuated Inversion Recovery (t2f), and T2-Weighted (t2w) [24]. These scans, depicted in Figure 2, capture the same anatomical structures using different imaging protocols, each providing distinct tissue contrasts. For instance, t1c enhances tumour visibility, while t2w is highly effective in defining fluid-filled regions. This variety is crucial for segmentation, as it provides complementary perspectives of brain anatomy, enabling the model to develop a more comprehensive understanding of the tumour region.

The original dataset employs a four-class segmentation system (0: Background/Healthy Tissue, 1: Non-Enhancing Tumour Core, 2: Enhancing Tumour, 3: Surrounding FLAIR Hyperintensity). For our study, we adopted a simplified binary encoding (0: Non-Tumour, 1: Intracranial Meningioma). This simplification, visualised in Figure 3, streamlined the design of our segmentation process and allowed our models to focus on the primary objective of identifying and delineating the whole tumour region.

### 4.2 Pre-processing

The BraTS dataset underwent initial pre-processing by the challenge organisers, including DICOM to NIfTI format (.nii.gz) conversion, modality co-registration to the SRI24 atlas space, and isotropic resampling to 1 mm³ voxel size. Standard skull-stripping was performed to anonymise the samples by removing non-brain tissues and protecting patient privacy [7, 24, 25].

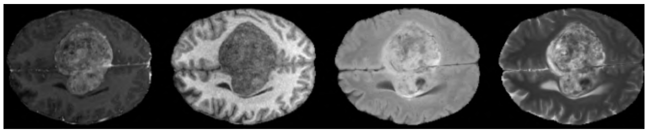


**Figure 2: The 4 different scans in our dataset per sample: t1n, t1c, t2f, and t2w [24].**

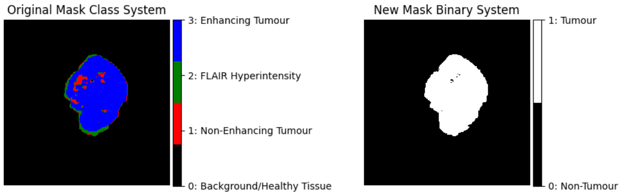


**Figure 3: Original Class Segmentation Mask (left) and the corresponding Binary Mask (right). The Four-Class System distinguishes between four types of regions: background/healthy tissue (black), non-enhancing tumour core (yellow), enhancing tumour (red), and surrounding FLAIR hyperintensity (green). The binary mask simplifies classification to just two categories: non-tumour (black) and tumour (white) [24]**

Additional pre-processing involved center-cropping to prioritise the brain region, reducing the inclusion of irrelevant background information. We then applied Z-Score Normalisation and Rescaling using sci-kit-learn's Exposure library to standardise intensity values across scans and mitigate any potential biases introduced by MRI machine calibration variations.

Since SAM operates on 2D RGB images, we converted grayscale slices to RGB, resizing as needed, and processed each slice individually due to SAM's 2D nature.

As a final step to prompt SAM for accurate segmentation, we generated a bounding box for each slice, as seen in Algorithm 1. This bounding box was defined by calculating the smallest rectangular region that completely encloses the entire tumour volume in the ground truth segmentation. For multiple disjoint tumours, the bounding box included all tumour regions, ensuring SAM received a comprehensive region for segmentation. This approach facilitated accurate tumour delineation across all MRI modalities. However, with this design, large bounding boxes could occur with very distant disjoint tumours, suggesting future research into multiple prompt inputs, where each disjoint tumour would have its own bounding box.

The design decision to compare the entire ground truth segmentation with SAM's predicted segmentation mask, assembled from per-slice predictions, was crucial. One might argue that comparing only the bounding box portion with the corresponding section in the ground truth is more accurate. However, SAM operates with a degree of human error tolerance, with a tendency to predict regions outside the bounding box prompt. Therefore, comparing the entire

**Algorithm 1: Bounding Box Generation**

```
Input: $seg\_mask$: Segmentation mask, $m$: Margin
Output: Bounding box coordinates or None
$contours \leftarrow \text{FindContours}(seg\_mask)$;
if $contours$ is empty then
    return None;
end
$(x_{min}, y_{min}, x_{max}, y_{max}) \leftarrow contours[0]$;
foreach $contour$ in $contours[1:]$ do
    $(x, y, w, h) \leftarrow \text{BoundingRect}(contour)$;
    $x_{min} \leftarrow \min(x_{min}, x)$;
    $y_{min} \leftarrow \min(y_{min}, y)$;
    $x_{max} \leftarrow \max(x_{max}, x + w)$;
    $y_{max} \leftarrow \max(y_{max}, y + h)$;
end
$x_{min} \leftarrow \max(0, x_{min} - m)$;
$y_{min} \leftarrow \max(0, y_{min} - m)$;
$x_{max} \leftarrow \min(\text{width of } seg\_mask, x_{max} + m)$;
$y_{max} \leftarrow \min(\text{height of } seg\_mask, y_{max} + m)$;
return $(x_{min}, y_{min}, x_{max}, y_{max})$;
```

3D ground truth segmentation mask with SAM's predicted segmentation mask was necessary, and straightforward as they share the same dimensional shape.

The dataset, consisting of 1000 samples (4000 scans), was split into training (68%), validation (20%), and testing (12%) sets, adhering to standard practices in medical image analysis. However, the official BraTS 2023 testing set was unavailable due to competition restrictions; hence, our test set was derived from the available training data.

## 4.3 Handling 3D to 2D

*4.3.1* ***Multi-Scan Per Sample.*** Each MRI modality was processed independently by SAM, generating separate predictions. The bounding box prompt, derived from the ground truth segmentation, was consistently applied across all modalities.

The final predicted mask for each sample was obtained by combining these predictions using a union operation, ensuring maximal tumour coverage and leveraging the unique information provided by each scan type. This approach aligns with multi-modal fusion techniques in medical imaging, where integrating different modalities enhances segmentation accuracy.

*4.3.2* ***Multi-Slice Per Sample.*** During segmentation, each sample's 3D MRI volume was processed slice-by-slice in the axial plane, chosen for its clinical relevance and ease of implementation. This plane is also the most intuitive from a non-medical perspective, as shown in Figure 4.

During validation and testing, each slice was processed individually, with the slice-by-slice results combined to produce a final 3D segmentation mask.

During the training phase, processing each slice resulted in significant time and memory limitations. This issue likely stemmed from SAM's input image enlargements during unwrapping and persisted despite meticulous memory management efforts. To mitigate this, we explored various slice-selection strategies. Initially, we employed a central slice method, choosing the slice located at the midpoint of each MRI sample. However, we found that this approach reduced the amount of usable data, as not all central slices contained tumours. Only 386 out of 1000 samples were viable.

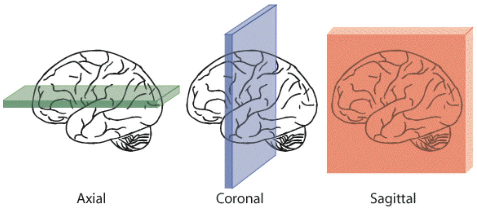


**Figure 4: The different planes of the brain [8].**

We then explored the max slice approach, which targeted the slice with the largest cross-sectional area of the tumour. We hypothesised that this would improve performance on large tumour instances. However, training on a single slice did not enhance segmentation accuracy because SAM lacked the information to generalise effectively, resulting in suboptimal learning outcomes.

We also considered a "4 Slice" Method, selecting slices representing 100%, 75%, 50%, and 25% of the tumour's maximum area. However, concerns about potential bias led us to reject this approach, as it likely would not generalise well across diverse tumour presentations.

Ultimately, we opted to train on every second slice per sample, which balanced training time and memory use. This completely avoided memory constraints, while still providing SAM with a diverse set of tumour examples, leading to better overall learning.

## 4.4 PEFT-SAM Implementation

Our Parameter-Efficient Fine-Tuning (PEFT) approach involved fine-tuning only the mask decoder of SAM, allowing it to adapt to the specific characteristics of our MRI data while minimising the risk of overfitting. This process was implemented using PyTorch [34].

We employed an exponentially decaying learning rate schedule and the Adam optimiser, known for its efficiency with sparse gradients, to manage the training process effectively.

Hyperparameter tuning was conducted using Optuna [1], systematically exploring the impact of various initial learning rates and optimisers, among other parameters. Initially, Dice Loss was used as the objective function; however, instability during training prompted a switch to a weighted combination of Mean-Squared Error (MSE) and Cross-Entropy Loss for more stable convergence.

We evaluated various loss computation strategies across four MRI modalities, including summing, maximum, and minimum methods. Ultimately, we made the design decision to compute the loss on the unionised 3D predicted segmentation mask, which incorporated all available information per sample, as this approach yielded the best results.

In addition, data shuffling and loaders were meticulously designed to ensure a balanced distribution of samples across training epochs. We initially introduced a difficulty-based categorisation of tumours to balance sample difficulty per epoch, inspired by curriculum learning. However, this measure was found to be redundant as tumour difficulty averaged out over training.

We opted to use SAM's single-output, despite its reported lower performance compared to the multi-output approach [23]. This design decision was made to avoid the risk of SAM consistently segmenting only the most visible part of the tumour, which is often the core. We aimed to focus on segmenting the entire tumour, not just its most apparent part, as using the most confident mask could limit accuracy and fail to provide a representative segmentation. Preliminary checks confirmed that SAM's single-output with point prompts, which is inherently ambiguous, led to incomplete tumour segmentation. Thus, we employed bounding boxes to reduce ambiguity and better capture the entire meningioma. Future research could explore the performance benefits of SAM's multi-output approach.

## 4.5 Framework Modification

To address the Research Question of whether combining different segmentation frameworks can improve performance, we integrated SAM with a 3D U-Net architecture. This hybrid approach combines SAM's Vision Transformer (ViT) [10], known for its robust feature extraction in 2D, with U-Net's CNN-based architecture [5, 39], which excels at capturing detailed 3D information. By merging SAM's efficient 2D segmentation capabilities with U-Net's advanced 3D volumetric processing, we aimed to harness the strengths of both frameworks to enhance overall segmentation performance.

For our U-Net architecture [39], we utilised a model inspired by Futrega et al.'s (NVIDIA) Optimised U-Net [12], which differs from Vanilla U-Net by incorporating deep supervision with additional decoder levels closer to the output, enhancing gradient flow and potentially improving segmentation accuracy.

While PEFT-SAM was trained slice-by-slice, U-Net processed full 3D volumes, leveraging the spatial context of the entire volume. Unlike SAM, which generates consistent outputs for the same input, U-Net predictions can vary. U-Net was trained with the same hyperparameters and loss function strategy as described in Section 4.4, except all parameters were updated during training. This differs from PEFT-SAM, where only the mask decoder was fine-tuned.

A significant difference between the architectures is the prompting method. SAM has a natural advantage by being limited to predict only on slices within the tumour volume due to the bounding box prompt. To ensure a fair comparison, we adjusted the U-Net segmentation process to consider the tumour volume's start and end points within the axial plane, emulating the bounding box prompt used by SAM. This adjustment ensured a fairer comparison, although future research should investigate the potential of prompting U-Net with a bounding box as an additional input channel.

*4.5.1* ***U-SAM-Combo.*** As illustrated in Figure 5, U-SAM-Combo was designed to harness the complementary strengths of PEFT-SAM and 3D U-Net. The input MRI was independently processed by both models, and their binary mask outputs were merged to produce the final segmentation mask. We explored three combination strategies.

First, the *union* approach, where the final segmentation mask is the union of PEFT-SAM and U-Net output. This maximised coverage by simply merging the predictions of both models.

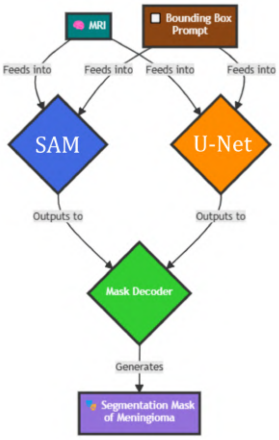


**Figure 5: Architecture of U-SAM-Combo.**

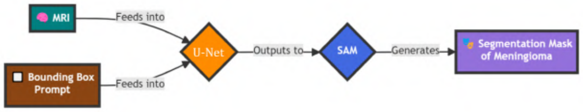


**Figure 6: Architecture of U1-SAM2.**

Secondly, the *intersection* strategy ensured precise segmentation by focusing only on regions where both model's predictions were in agreement.

Lastly, we investigated a *weighted* method. This required a preliminary accuracy evaluation of each model's performance with the ground truth segmentation mask to determine the relative performance of PEFT-SAM and U-Net per sample. Thus, we acknowledge that this approach is more research-focused and does not yet have immediate clinical applicability.

*4.5.2* ***U1-SAM2.*** As shown in Figure 6, the U1-SAM2 architecture employs a sequential approach where the MRI is first processed by U-Net to generate an initial prediction. This prediction then serves as an input, guidance mask for PEFT-SAM, refining the segmentation based on U-Net's output. The final prediction from PEFT-SAM, guided by the U-Net mask, is used for evaluation. This approach combines SAM's precise segmentation capabilities with the broader contextual understanding provided by U-Net

To achieve this, SAM's `mask_input` parameter in its `predict` method was utilised for integration [23]. Although under-documented, this feature was crucial for combining U-Net's output with SAM. Notably, SAM expects input masks containing unthresholded logits (floating-point confidence values), not binary masks. Tools from MicroSAM [2] were adapted to convert U-Net's binary output into logits compatible with SAM, facilitating seamless integration of the two models.

## 5 RESULTS AND DISCUSSION

The Dice Similarity Coefficient (DSC) was chosen as our primary evaluation metric due to its effectiveness in measuring the accuracy of tumour delineation, which is crucial given the variable size and shape of tumours. The DSC quantifies the overlap between the predicted segmentation and the ground truth, offering a robust measure of segmentation performance. The formula used is:

$$\mathrm{DSC} = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{1}$$

where A is the predicted segmentation and B is the ground truth. The DSC ranges from 0 (no overlap) to 1 (perfect overlap).

For our evaluations, we utilised MONAI's implementation of the DiceLoss metric, which provides a reliable and standardised approach for calculating the DSC and is widely employed in medical imaging applications [4]

### 5.1 Vanilla SAM Baseline

Our evaluation of Vanilla SAM yielded an overall DSC of approximately 0.84 across various MRI modalities (Table 1). This consistency across modalities, particularly with t1c images achieving the highest score of 0.851, highlights SAM's robustness despite being primarily trained on non-medical data.

The visualisations in Appendix A, Figure 12, reinforce these findings. Despite the model's limited training on medical data, the outputs for each scan type demonstrate that Vanilla SAM performs well, with predictions generally aligning closely with the ground truth. This observation is particularly evident in the t1c images, where the clear visibility of tumours due to contrast enhancement allows SAM to leverage its pre-trained capabilities effectively. Conversely, the performance dips in the t2w scans, where the less pronounced tumour visibility challenges the model's ability to make accurate predictions. This variation in performance across modalities is consistent with the visual feedback, highlighting the model's strengths and weaknesses in different imaging contexts.

This baseline result was unexpectedly high, given our assumption and prior research that SAM, primarily trained on non-medical data, would struggle with the unique challenges presented by medical imaging. This suggests that SAM's pre-trained weights have considerable potential for zero-shot learning, even in the specialised domain of MRI imaging. However, while the results are promising, the model's performance still falls short of the near-perfect accuracy required for clinical applications, indicating a need for further fine-tuning.

**Table 1: Vanilla SAM Testing Dice Scores across different MRI modalities. As seen by the bold value, t1c achieved the highest performance, which is expected due to the increased visibility of tumours in contrast-enhanced T1-weighted images. However, the overall score remains most relevant for generalisation.**

| t1c | t1n | t2f | t2w | Overall |
|---|---|---|---|---|
| **0.851** | 0.809 | 0.834 | 0.807 | 0.841 |

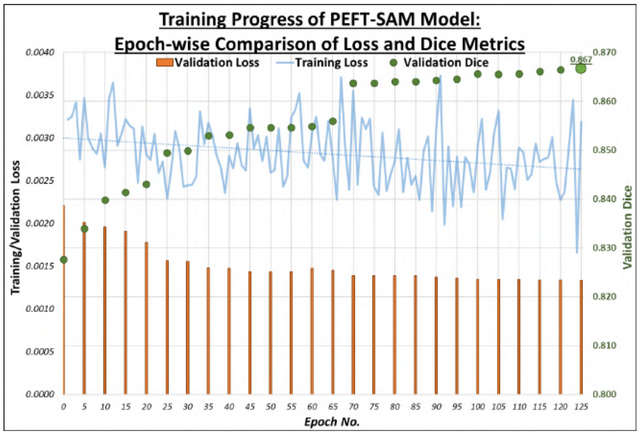


**Figure 7: PEFT-SAM Training Progression, showing the relationship between Training Loss, Validation Loss, and Validation Dice across 125 epochs. The training loss is represented by a blue line, validation loss by orange bars, and validation Dice score by green markers.**

### 5.2 PEFT-SAM

Our initial attempts to fine-tune SAM were hindered by underestimating the computational demands, which led to signs of underfitting and no significant gains in segmentation accuracy despite extensive training. After tweaking the training scheme to the approach outlined in Section 4.4, the following results were observed:

In Figure 7, the training loss, although somewhat unstable and challenging to interpret for clear convergence, shows improvement from the initial training with the Dice Loss function. Despite these improvements, our focus remains on enhancing SAM's overall DSC score, as observed in the graph.

However, this came with a trade-off: the performance gains were not uniform across all scan types. Specifically, the successful overall improvement in DSC from our training was accompanied by a deterioration in per-scan type performance. This can be seen in Figure 8.

Appendix A, Figure 13 provides a detailed visualisation of PEFT-SAM's performance across different scan types that reinforces this. While the overall accuracy improved relative to Vanilla, the visuals highlight notable variations in performance between scan types. The deterioration of some modalities reflects the challenge of achieving consistent accuracy. The visualisation suggests that our PEFT-SAM, despite its overall improvement from tuning, now struggles to maintain consistent performance across diverse scan types. This result underscores the inherent complexity in tuning models for diverse medical imaging data.

We attribute these variations to our model's enhanced ability to better identify different tumour features depending on scan type, which varied in visibility. The model effectively became a combination of "weak" classifiers, each specialised in certain aspects of the tumour per scan type, leading to a stronger overall classifier for the four scan types used. However, this dependency also limits the generalisation of PEFT-SAM to other, unseen types of medical data.

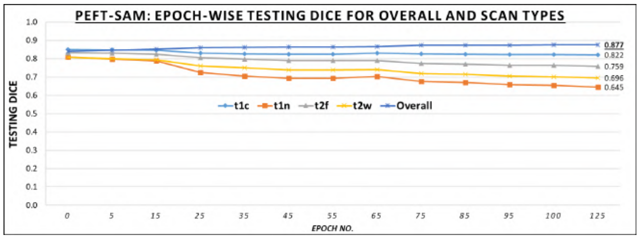


**Figure 8: PEFT-SAM Testing Dice for different scan types and overall across epochs. While overall performance improves consistently, some individual scans exhibit deterioration in Dice scores, highlighting the challenges in achieving uniform accuracy across different scan types.**

While we had anticipated more significant improvement, the overall 3.6% gain still represents a meaningful enhancement over the baseline. Although the per-scan type performance deterioration was concerning, this progress was crucial for establishing a solid foundation for our framework modification to SAM and offered valuable insights into the complexities of fine-tuning SAM for medical imaging tasks. Given these findings, we recommend a more nuanced approach to model tuning, considering scan-specific characteristics and potentially employing advanced techniques to address the variability in performance. Due to time constraints, any further fine-tuning breakthroughs were curtailed, and focus was shifted towards the primary objective of our project: implementing and evaluating our framework modification on SAM.

## 5.3 3D U-Net

Training the 3D U-Net model yielded significant improvements, with the DSC rising from 0.60 to 0.82. This performance is comparable to our Vanilla SAM model of 0.84 and underscores U-Net's capability to handle 3D volumes and multiple channels effectively. The progression of training is shown in Figure 9, demonstrating the learning curve of the model.

Figure 10 reveals insightful contrasts between 3D U-Net and PEFT-SAM. For example, the t1n scan type, which posed significant challenges for Vanilla SAM (one of the least performant, as seen in Table 1), also proved difficult for U-Net when given just that scan type. This underscores the inherent challenges associated with certain scan types and the limitations of both models in addressing these challenges.

However, it is important to note that the comparison may not be entirely fair. U-Net was trained on the full 3D volume with four channels for each scan type, which differs significantly from PEFT-SAM's per-scan type predictor approach. This offers U-Net a more comprehensive context, not necessarily to predict per scan types.

Appendix B provides visualisations that show how U-Net's ability to process the entire 3D volume with multiple channels allows it to achieve precise edge delineation, especially in complex regions of the tumour. This is evident from the detailed contours and boundaries that U-Net can delineate, which are often more refined than those produced by Vanilla SAM or PEFT-SAM. This precision is critical in medical imaging, where accurate tumour delineation can significantly impact diagnostic and treatment decisions.

The introduction of bounding box emulation into the U-Net framework led to observable improvements in DSC, specifically a 1.2% accuracy gain. This enhancement aligns with expectations, as the bounding box provides additional contextual information closer to what SAM uses. However, it was not as substantial as anticipated. We hypothesise that with full bounding box emulation, where the U-Net has access to all the additional context provided by the bounding box that SAM is given, further gains could be achieved. Future research should focus on fully integrating bounding box emulation and exploring additional improvements to enhance generalisation across various scan types.

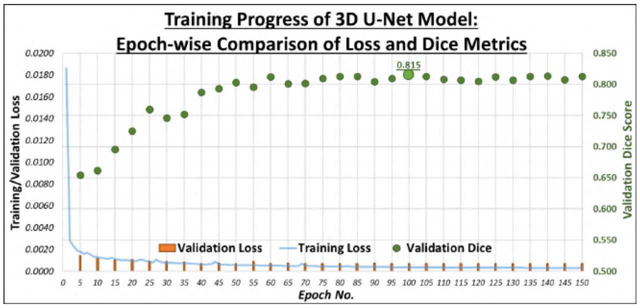


**Figure 9: Training Progression of 3D U-Net, showing the relationship between Training Loss, Validation Loss, and Validation Dice across 125 epochs. The training loss is represented by a blue line, validation loss by orange bars, and validation Dice score by green markers.**

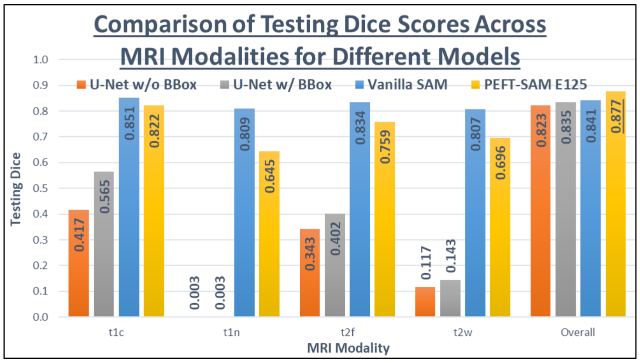


**Figure 10: Testing Dice comparison across different MRI modalities and the overall for four models: U-Net without BBox, U-Net with BBox, Vanilla SAM, and PEFT-SAM E125. Results highlight the variation in performance between the models for each modality, emphasising the effectiveness of each approach in different scenarios. The PEFT-SAM E125 model demonstrates exceptional overall performance, whereas the Vanilla SAM excels in certain specific modalities.**

## 5.4 U-SAM-Combo

The U-SAM-Combo approach demonstrates a mixed picture of the effectiveness of combining the U-Net and SAM predictions, with each combination strategy (Union, Intersection, and Weighted) offering distinct strengths and limitations.

*5.4.1* ***Union.*** The Union method, which aggregates predictions by taking the union of U-Net and SAM outputs, consistently improved upon using U-Net alone. It particularly excelled when U-Net and SAM predictions align, suggesting that when both models agree, it effectively consolidates their agreement for comprehensive tumour segmentation. Notably, the peak Dice score achieved with bounding box emulation in U-Net was 0.886 with PEFT-SAM E125, as detailed in Table 2.

Visualisations in Appendix C illustrate the Union method's effectiveness in leveraging both models' strengths, enhancing overall coverage. This improvement is especially evident with higher training epochs of SAM, indicating that the Union method benefits from well-trained models. Thus, it proves to be a robust strategy in scenarios where SAM predictions are reliable.

*5.4.2* ***Intersection.*** The Intersection method, which focuses solely on the overlap between U-Net and SAM predictions, did not perform as anticipated. This approach, designed to minimise false positives by excluding areas of disagreement, instead exacerbated errors. able 2 shows consistently lower Dice scores with the Intersection method versus the Union method.

Appendix C provides visuals that highlight the limitations of the Intersection method, particularly in failing to capture significant tumour portions. This shortfall occurs because errors from either model result in a substantial loss of true positives. The minimal impact of bounding box emulation further suggests that the Intersection method is less effective in incorporating U-Net's enhancements, further underscoring the need for caution due to its tendency to reduce the overall predicted area.

*5.4.3* ***Weighted.*** The Weighted method, which integrates predictions based on model performance alignment, yielded the best overall results. With Dice scores peaking at 0.918, as shown in Table 2, this method effectively balances the contributions of both models, particularly excelling in handling edge cases where other methods fall short.

Appendix C provides visual evidence of the Weighted method's superior performance. Its adaptive nature facilitates a more nuanced integration of predictions, enhancing segmentation accuracy and detail, particularly in complex cases where individual model predictions are inconsistent. However, it is important to note again that the requirement for intermediate evaluation of U-Net and SAM predictions introduces potential bias. Despite this, the superior performance of the Weighted method suggests it is a promising strategy for combining segmentation outputs.

## 5.5 U1-SAM2

The U1-SAM2 model represents an effort to enhance SAM's predictions by integrating insights from U-Net, focusing on improving segmentation accuracy. Although it shows promise, the results reveal several challenges.

**Table 2: U-SAM-Combo Testing Dice Scores for different PEFT-SAM versions and combination strategies, utilising the most performant U-Net (E100). Raw U-Net and SAM Dice scores are included for comparison. The bold values in each row indicate the best performance for each configuration.**

| U-Net w/ BBox | SAM Version | U-Net Only | SAM Only | Union | Intersection | Weighted |
|---|---|---|---|---|---|---|
| False | Vanilla | 0.823 | 0.841 | 0.811 | 0.841 | **0.908** |
| True | Vanilla | 0.835 | 0.841 | 0.837 | 0.841 | **0.911** |
| False | E65 | 0.823 | 0.867 | 0.845 | 0.831 | **0.912** |
| True | E65 | 0.835 | 0.867 | 0.871 | 0.831 | **0.915** |
| False | E100 | 0.823 | 0.876 | 0.858 | 0.826 | **0.915** |
| True | E100 | 0.835 | 0.875 | 0.884 | 0.826 | **0.917** |
| False | E125 | 0.823 | 0.877 | 0.860 | 0.826 | **0.916** |
| True | E125 | 0.835 | 0.877 | 0.886 | 0.826 | **0.918** |

**Table 3: U1-SAM2 Testing Dice Scores for different PEFT-SAM versions with the most performant U-Net (E100). Raw U-Net and SAM Dice scores are included for comparison. The bold values in each row indicate the best performance for each configuration.**

| U-Net w/ BBox | SAM Version | U-Net Only | SAM Only | U1-SAM2 |
|---|---|---|---|---|
| False | Vanilla | 0.823 | **0.841** | 0.837 |
| True | Vanilla | 0.835 | **0.841** | 0.837 |
| False | E65 | 0.823 | **0.867** | 0.835 |
| True | E65 | 0.835 | **0.867** | 0.835 |
| False | E100 | 0.823 | **0.875** | 0.847 |
| True | E100 | 0.835 | **0.875** | 0.847 |
| False | E125 | 0.823 | **0.877** | 0.851 |
| True | E125 | 0.835 | **0.877** | 0.851 |

While U1-SAM2 improved upon U-Net's raw performance in segmentation, particularly in challenging tasks, the final mask quality sometimes deteriorated compared to PEFT-SAM alone. This deterioration was most noticeable in cases where PEFT-SAM exhibited lower confidence, with U1-SAM2 introducing spurious "on" pixels that degraded the overall mask quality.

Appendix C visualisations reveal how U1-SAM2 integration results in noisy masks, especially in challenging regions where PEFT-SAM was more accurate. The introduction of these spurious pixels suggests that U1-SAM2 may be overly sensitive to U-Net's predictions, particularly when they diverge from SAM's output. This sensitivity leads to an over-correction, resulting in false positives that reduce the overall segmentation accuracy.

As shown in Table 3, results suggest that U1-SAM2 could enhance segmentation accuracy, especially when U-Net predictions offer beneficial additional insights. However, it still needs further refinement. The model's tendency to introduce noise suggests the need for more sophisticated combination strategies that can better discern when to trust U-Net's predictions and when to rely more heavily on SAM.

## 6 SUMMARY OF MAIN FINDINGS

This study evaluated the performance of various Deep Learning models for MRI intracranial meningioma segmentation, including SAM, PEFT-SAM, 3D U-Net, U-SAM-Combo, and U1-SAM2.

**Vanilla SAM:** Despite being pre-trained on non-medical data, SAM achieved an overall DSC of 0.841, demonstrating its potential for zero-shot learning in medical imaging.

**PEFT-SAM:** Fine-tuning SAM with parameter-efficient techniques resulted in a modest DSC improvement of approximately 3.6%. This gain came at the cost of uneven performance across different scan types, suggesting that while overall segmentation accuracy improved, the fine-tuned model struggled with maintaining uniformity across diverse imaging conditions.

**3D U-Net:** Training a 3D U-Net from scratch achieved a DSC of 0.835, comparable to Vanilla SAM. Marginal improvements were observed with bounding box emulation, suggesting potential for further enhancement.

**U-SAM-Combo:** Combining U-Net and SAM outputs using various strategies revealed that the Weighted method outperformed others with a peak DSC of 0.918. This method effectively integrated the strengths of both models, particularly in challenging cases, though it introduced complexity and potential bias through intermediate evaluations.

**U1-SAM2:** U1-SAM2 attempted to refine SAM's predictions with U-Net insights. While it improved upon U-Net's raw performance, it sometimes degraded mask quality by introducing noise and spurious pixels, particularly when PEFT-SAM's confidence was low. This suggests that this model may be overly sensitive to U-Net's predictions, highlighting the need for more strategies that better balance model contributions.

These findings, summarised in Table 4, underscore the promise of integrating and fine-tuning advanced Deep Learning models for medical image segmentation. The U-SAM-Combo approach, particularly the Weighted method, shows significant promise, though careful consideration of model alignment and bias mitigation is necessary.

**Table 4: Comparison of Overall Testing Dice Scores across Various Models. As seen by the bolded value, the U-SAM-Combo Weighted model achieves the highest performance with a Dice of 0.918, significantly outperforming the baseline Vanilla SAM model. This improvement is likely due to the weighted combination approach that leverages the strengths of both U-Net and SAM architectures.**

| Model | Testing Dice |
|---|---|
| Vanilla SAM | 0.841 |
| U-Net w/o BBox | 0.823 |
| U-Net w/ BBox | 0.835 |
| U-SAM-Combo *(Union)* | 0.886 |
| U-SAM-Combo *(Intersection)* | 0.826 |
| U-SAM-Combo *(Weighted)* | **0.918** |
| U1-SAM2 | 0.851 |

## 7 CONCLUSIONS AND FUTURE WORK

In this study, we aimed to enhance the Segment Anything Model's (SAM) segmentation performance of intracranial meningiomas in 3D medical imaging, through fine-tuning and modifying its framework. Our primary research question was whether integrating SAM with U-Net could improve segmentation accuracy for intracranial meningiomas. We addressed this question by employing two novel approaches, U-SAM-Combo, which combines U-Net and SAM predictions through union, intersection, and weighted combination methods, and U1-SAM2, which feeds U-Net's prediction into SAM as an additional prompt.

Our most notable achievement reveals that the U-SAM-Combo approach achieved improvements of 4.5% and 7.7% in accuracy over the baseline SAM model, respectively. This enhancement demonstrates the potential of integrating SAM into medical imaging workflows and contributes valuable insights to the medical SAM community.

However, we encountered several challenges that highlighted the need for more substantial computational resources and extended training periods to achieve significant improvements. In addition, SAM exhibited limitations in handling huge tumours, where fine-tuning and adaptations yielded only marginal improvements.

The U-SAM-Combo approach showed considerable promise, confirming that combining U-Net and SAM predictions can enhance segmentation performance. However, the U1-SAM2 model, while innovative, introduced spurious pixels in certain cases, particularly when SAM's confidence was low, indicating that this method requires further refinement.

These results, although modest, underscore SAM's potential in medical imaging, and contribute to the growing body of knowledge within the medical SAM community. While SAM's capabilities in intracranial meningioma segmentation were not fully realised in this research, our work indicates that with more training and optimisation, SAM could become a valuable tool in medical imaging applications.

Future work should focus on several key areas to build upon these findings. Enhancing the full integration of SAM's bounding box prompts into U-Net training may yield more robust segmentation results. In addition, exploring the reverse of our U1-SAM2 approach, U2-SAM1, where SAM's predictions are used as input guidance masks for U-Net, could also improve accuracy by leveraging the strengths of both models in a complementary manner. Finally, investigating alternative prompting methods, such as point-based or text-based prompts, may provide further insights into optimising SAM for various medical imaging tasks.

In conclusion, while our study highlights the promise of SAM for medical imaging applications, further research, resources, and collaboration are necessary to fully realise its potential and pave the way for the model's broader application in clinical settings.

## 8 ETHICAL, PROFESSIONAL, AND INTELLECTUAL PROPERTY

All data used in this study has been anonymised to ensure privacy. Despite this, ethical approval was obtained from the Inter-Faculty Research Ethics Committee due to the use of medical data, ensuring that no patients are identifiable. Additionally, no patented methods or libraries were utilised. All external code libraries employed were sourced from open-source platforms, and any adapted code has been properly cited.

## 9 ACKNOWLEDGEMENTS

We extend our heartfelt thanks to Patrick Marais and Fred Nicolls for their invaluable guidance and support throughout this academic endeavour. We also appreciate the BraTS team for their dataset, which was crucial to this study. Finally, we acknowledge the University of Cape Town's ICTS High-Performance Computing team [31] for providing the A100 GPU Core Cluster, which facilitated our training and evaluations.

## REFERENCES

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2623–2631.

[2] Anwai Archit, Sushmita Nair, Nabeel Khalid, Paul Hilt, Vikas Rajashekar, Marei Freitag, Sagnik Gupta, Andreas Dengel, Sheraz Ahmed, and Constantin Pape. 2023. Segment Anything for Microscopy. *bioRxiv* (2023). https://doi.org/10.1101/2023.08.21.554208

[3] Nhat-Tan Bui, Dinh-Hieu Hoang, Minh-Triet Tran, Gianfranco Doretto, Donald Adjeroh, Brijesh Patel, Arabinda Choudhary, and Ngan Le. 2024. SAM3D: Segment Anything Model in Volumetric Medical Images. arXiv:2309.03493 [eess.IV] https://arxiv.org/abs/2309.03493

[4] M. Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, Vishwesh Nath, Yufan He, Ziyue Xu, Ali Hatamizadeh, Andriy Myronenko, Wentao Zhu, Yun Liu, Mingxin Zheng, Yucheng Tang, Isaac Yang, Michael Zephyr, Behrooz Hashemian, Sachidanand Alle, Mohammad Zalbagi Darestani, Charlie Budd, Marc Modat, Tom Vercauteren, Guotai Wang, Yiwen Li, Yipeng Hu, Yunguan Fu, Benjamin Gorman, Hans Johnson, Brad Genereaux, Barbaros S. Erdal, Vikash Gupta, Andres Diaz-Pinto, Andre Dourson, Lena Maier-Hein, Paul F. Jaeger, Michael Baumgartner, Jayashree Kalpathy-Cramer, Mona Flores, Justin Kirby, Lee A. D. Cooper, Holger R. Roth, Daguang Xu, David Bericat, Ralf Floca, S. Kevin Zhou, Haris Shuaib, Keyvan Farahani, Klaus H. Maier-Hein, Stephen Aylward, Prerna Dogra, Sebastien Ourselin, and Andrew Feng. 2022. MONAI: An open-source framework for deep learning in healthcare. arXiv:2211.02701 [cs.LG] https://arxiv.org/abs/2211.02701

[5] Rahul Chauhan, Kamal Kumar Ghanshala, and RC Joshi. 2018. Convolutional neural network (CNN) for image detection and recognition. In *2018 first international conference on secure cyber computing and communication (ICSCCC)*. IEEE, 278–282.

[6] Cheng Chen, Juzheng Miao, Dufan Wu, Zhiling Yan, Sekeun Kim, Jiang Hu, Aoxiao Zhong, Zhengliang Liu, Lichao Sun, Xiang Li, Tianming Liu, Pheng-Ann Heng, and Quanzheng Li. 2023. MA-SAM: Modality-agnostic SAM Adaptation for 3D Medical Image Segmentation. arXiv:2309.08842 [cs.CV] https://arxiv.org/abs/2309.08842

[7] Mildred Cho. 2021. Rising to the challenge of bias in health care AI. *Nature Medicine* 27 (12 2021), 1–2. https://doi.org/10.1038/s41591-021-01577-2

[8] Stuart Clare. 1997. Functional MRI : Methods and Applications. In *Functional MRI : Methods and Applications*. https://api.semanticscholar.org/CorpusID:8441926

[9] 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* abs/2304.04155 (2023). https://api.semanticscholar.org/CorpusID:258049163

[10] 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. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR* abs/2010.11929 (2020). arXiv:2010.11929 https://arxiv.org/abs/2010.11929

[11] Weijia Feng, Lingting Zhu, and Lequan Yu. 2023. Cheap Lunch for Medical Image Segmentation by Fine-tuning SAM on Few Exemplars. arXiv:2308.14133 [cs.CV] https://arxiv.org/abs/2308.14133

[12] Michał Futrega, Alexandre Milesi, Michał Marcinkiewicz, and Pablo Ribalta. 2021. Optimized U-Net for brain tumor segmentation. In *International MICCAI brainlesion workshop*. Springer, 15–29.

[13] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 2023. 3DSAM-adapter: Holistic Adaptation of SAM from 2D to 3D for Promptable Medical Image Segmentation. arXiv:2306.13465 [cs.CV] https://arxiv.org/abs/2306.13465

[14] J. A. Hartigan and M. A. Wong. 1979. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 1 (1979), 100–108. http://www.jstor.org/stable/2346830

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15979–15988. https://doi.org/10.1109/CVPR52688.2022.01553

[16] Chuanfei Hu and Xinde Li. 2023. When SAM Meets Medical Images: An Investigation of Segment Anything Model (SAM) on Multi-phase Liver Tumor Segmentation. *ArXiv* abs/2304.08506 (2023). https://api.semanticscholar.org/CorpusID:258187428

[17] Mingzhe Hu, Yuheng Li, and Xiaofeng Yang. 2023. SkinSAM: Empowering Skin Cancer Segmentation with Segment Anything Model. arXiv:2304.13973 [cs.CV] https://arxiv.org/abs/2304.13973

[18] Alicia Tosoni Vincenzo Di Nunno Lidia Gatto Raffaele Lodi Ilaria Maggio, Enrico Franceschi and Alba A Brandes. 2021. Meningioma: not always a benign tumor. A review of advances in the treatment of meningiomas. *CNS Oncology* 10, 2 (2021), CNS72. https://doi.org/10.2217/cns-2021-0003 arXiv:https://doi.org/10.2217/cns-2021-0003 PMID: 34015955.

[19] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein. 2018. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. arXiv:1809.10486 [cs.CV]

[20] Ashish Issac, M. Partha Sarathi, and Malay Kishore Dutta. 2015. An adaptive threshold based image processing technique for improved glaucoma detection and classification. *Computer Methods and Programs in Biomedicine* 122, 2 (2015), 229–244. https://doi.org/10.1016/j.cmpb.2015.08.002

[21] Wei Ji, Jingjing Li, Qi Bi, Wenbo Li, and Li Cheng. 2023. Segment Anything Is Not Always Perfect: An Investigation of SAM on Different Real-world Applications. *ArXiv* abs/2304.05750 (2023). https://api.semanticscholar.org/CorpusID:258079065

[22] Girish Katti, Syeda Arshiya Ara, and Ayesha Shireen. 2011. Magnetic resonance imaging (MRI)–A review. *International journal of dental clinics* 3, 1 (2011), 65–70.

[23] 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]

[24] 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]

[25] Karim Lekadira, Richard Osuala, Catherine Simone Gallin, Noussair Lazrak, Kaisar Kushibar, Gianna Tsakou, Susanna Auss'o, Leonor Cerd'a Alberich, Konstantinos Marias, Manolis Tskinakis, Sara Colantonio, Nickolas Papanikolaou, Zohaib Salahuddin, H. Woodruff, Philippe Lambin, and Luis Mart'i-Bonmat'i. 2021. FUTURE-AI: Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in Medical Imaging. *ArXiv* abs/2109.09658 (2021). https://api.semanticscholar.org/CorpusID:237572282

[26] Shengze Li, Jianjian Cao, Peng Ye, Yuhan Ding, Chongjun Tu, and Tao Chen. 2024. ClipSAM: CLIP and SAM Collaboration for Zero-Shot Anomaly Segmentation. arXiv:2401.12665 [cs.CV]

[27] Yuheng Li, Mingzhe Hu, and Xiaofeng Yang. 2023. Polyp-SAM: Transfer SAM for Polyp Segmentation. arXiv:2305.00293 [eess.IV] https://arxiv.org/abs/2305.00293

[28] Yunxiang Li, Bowen Jing, Zihan Li, Jing Wang, and You Zhang. 2023. nnSAM: Plug-and-play Segment Anything Model Improves nnUNet Performance. arXiv:2309.16967 [cs.CV]

[29] Xian Lin, Yangyang Xiang, Li Zhang, Xin Yang, Zengqiang Yan, and Li Yu. 2023. SAMUS: Adapting Segment Anything Model for Clinically-Friendly and Generalizable Ultrasound Image Segmentation. arXiv:2309.06824 [cs.CV]

[30] 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), 654. https://doi.org/10.1038/s41467-024-44824-z

[31] University of Cape Town, Timothy Carr, and Andrew Lewis. 2023. UCT HPC Facility. https://doi.org/10.5281/zenodo.10021613

[32] Christian Ogasawara, Brandon D. Philbrick, and D. Cory Adamson. 2021. Meningioma: A Review of Epidemiology, Pathology, Diagnosis, Treatment, and Future Directions. *Biomedicines* 9, 3 (2021). https://doi.org/10.3390/biomedicines9030319

[33] Jay N. Paranjape, Nithin Gopalakrishnan Nair, Shameema Sikder, S. Swaroop Vedula, and Vishal M. Patel. 2023. AdaptiveSAM: Towards Efficient Tuning of SAM for Surgical Scene Segmentation. arXiv:2308.03726 [cs.CV]

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[35] Moritz Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, David Robben, Alexandre Hutton, Tassilo Friedrich, Teresa Zarth, Johannes Bürkle, The Baran, Bjoern Menze, Gabriel Broocks, Lukas Meyer, and Jan Kirschke. 2022. ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific Data* 9 (12 2022), 762. https://doi.org/10.1038/s41597-022-01875-5

[36] Quan Quan, Fenghe Tang, Zikang Xu, Heqin Zhu, and S. Kevin Zhou. 2023. Slide-SAM: Medical SAM Meets Sliding Window. arXiv:2311.10121 [cs.CV]

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]

[38] Tianyi Ren, Ethan Honey, Harshitha Rebala, Abhishek Sharma, Agamdeep Chopra, and Mehmet Kurt. 2024. An Optimization Framework for Processing and Transfer Learning for the Brain Tumor Segmentation. *ArXiv* abs/2402.07008 (2024). https://api.semanticscholar.org/CorpusID:267626912

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597 [cs.CV]

[40] Lv Tang, Haoke Xiao, and Bo Li. 2023. Can SAM Segment Anything? When SAM Meets Camouflaged Object Detection. *ArXiv* abs/2304.04709 (2023). https://api.semanticscholar.org/CorpusID:258048579

[41] Jure Zbontar, Florian Knoll, Anuroop Sriram, Matthew Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael G. Rabbat, Pascal Vincent, James Pinkerton, Duo Wang, Nafissa Yakubova, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. 2018. fastMRI: An Open Dataset and Benchmarks for Accelerated MRI. *ArXiv* abs/1811.08839 (2018). https://api.semanticscholar.org/CorpusID:53759905

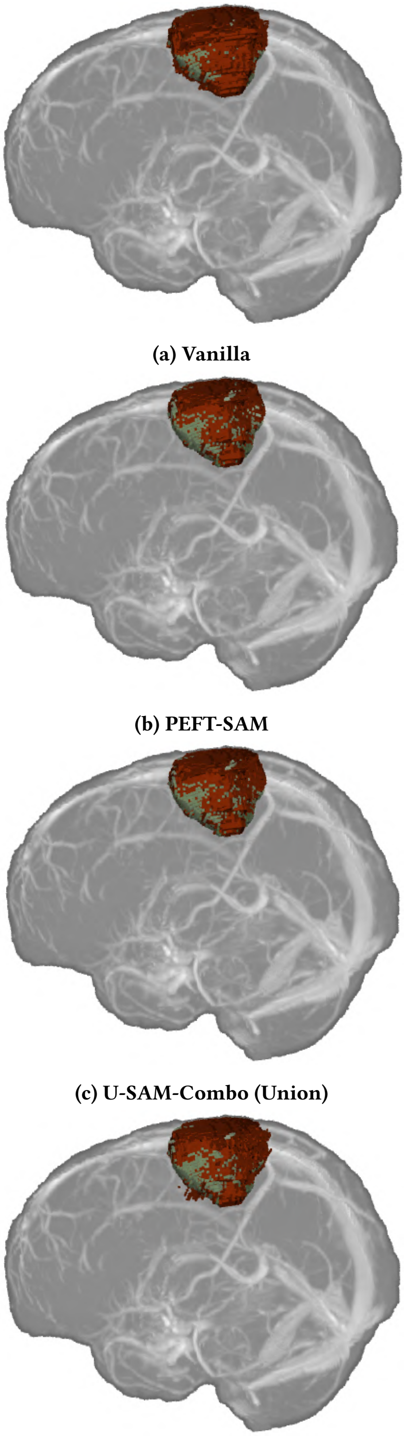
[42] Yichi Zhang, Zhenrong Shen, and Rushi Jiao. 2024. Segment anything model for medical image segmentation: Current applications and future directions. *Computers in Biology and Medicine* 171 (2024), 108238. https://doi.org/10.1016/j.compbiomed.2024.108238

# 10 SUPPLEMENTARY INFORMATION

## A VANILLA SAM TO PEFT-SAM VISUALS

*Note: The following visualisations use the slice with the largest cross-sectional area of the tumour. The colour-coding is as follows:*

- Green: Correctly identified tumour regions (True Positive)
- Red: Incorrectly identified tumour regions (False Positive)
- Blue: Missed tumour regions (False Negative)



**(a) Vanilla**

**(b) PEFT-SAM**

**(c) U-SAM-Combo (Union)**

**(d) U1-SAM2**

**Figure 11: A 3D Visualisation of our model's predictions on a sample for subjective assessment. Each image shows ground truth (green) and predictions (red). The amount of red indicates prediction accuracy: (a) Vanilla SAM, (b) PEFT-SAM, (c) U-SAM-Combo (Union), and (d) U1-SAM2. This highlights how our models have improved from Vanilla SAM as more red pixels are shaved off, but still imperfect.**

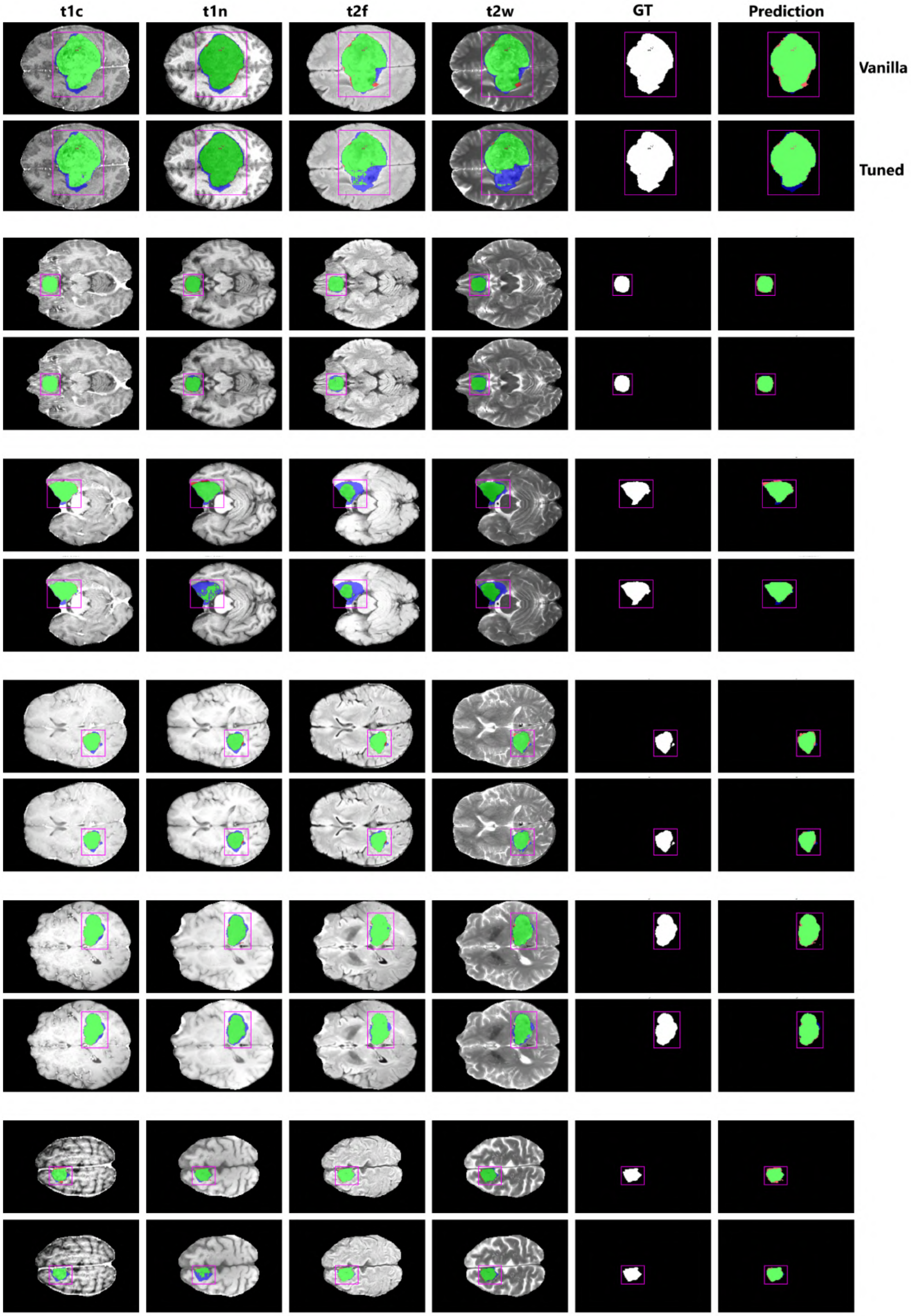


Figure 12: Comparison of Vanilla SAM and PEFT-SAM (E125) on good examples. The figure shows the per-scan predictions for each modality (t1c, t1n, t2f, t2w), the ground truth, and overall prediction, from left to right. In each grouping, the first row displays Vanilla SAM results, while the second row shows PEFT-SAM results. Although PEFT-SAM shows improved performance with fewer false positives, Vanilla SAM's predictions are already quite good in some cases, and tuning has a minimal effect where predictions are accurate.

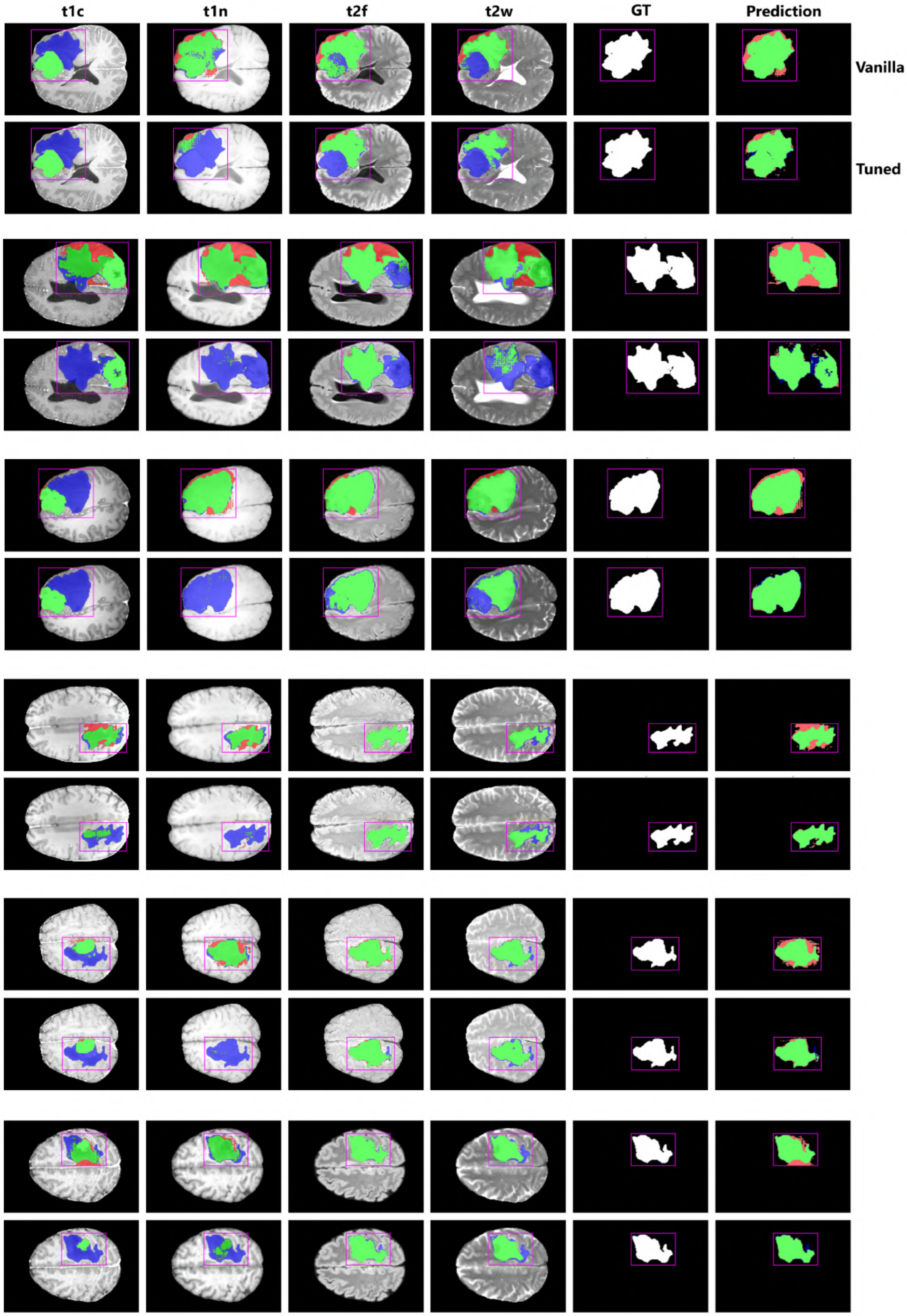


**Figure 13: Comparison of Vanilla SAM and PEFT-SAM (E125) on challenging examples. The figure shows the per-scan predictions for each modality (t1c, t1n, t2f, t2w), the ground truth, and overall prediction, from left to right. In each grouping, the first row displays Vanilla SAM results, while the second row shows PEFT-SAM results. PEFT-SAM shows improved results with fewer false positives, especially on complex edges and large tumours. However, the per-scan type deterioration is evident, especially t1n, though it is clear overall performance has improved relative to Vanilla SAM.**

## B 3D U-NET VISUALS

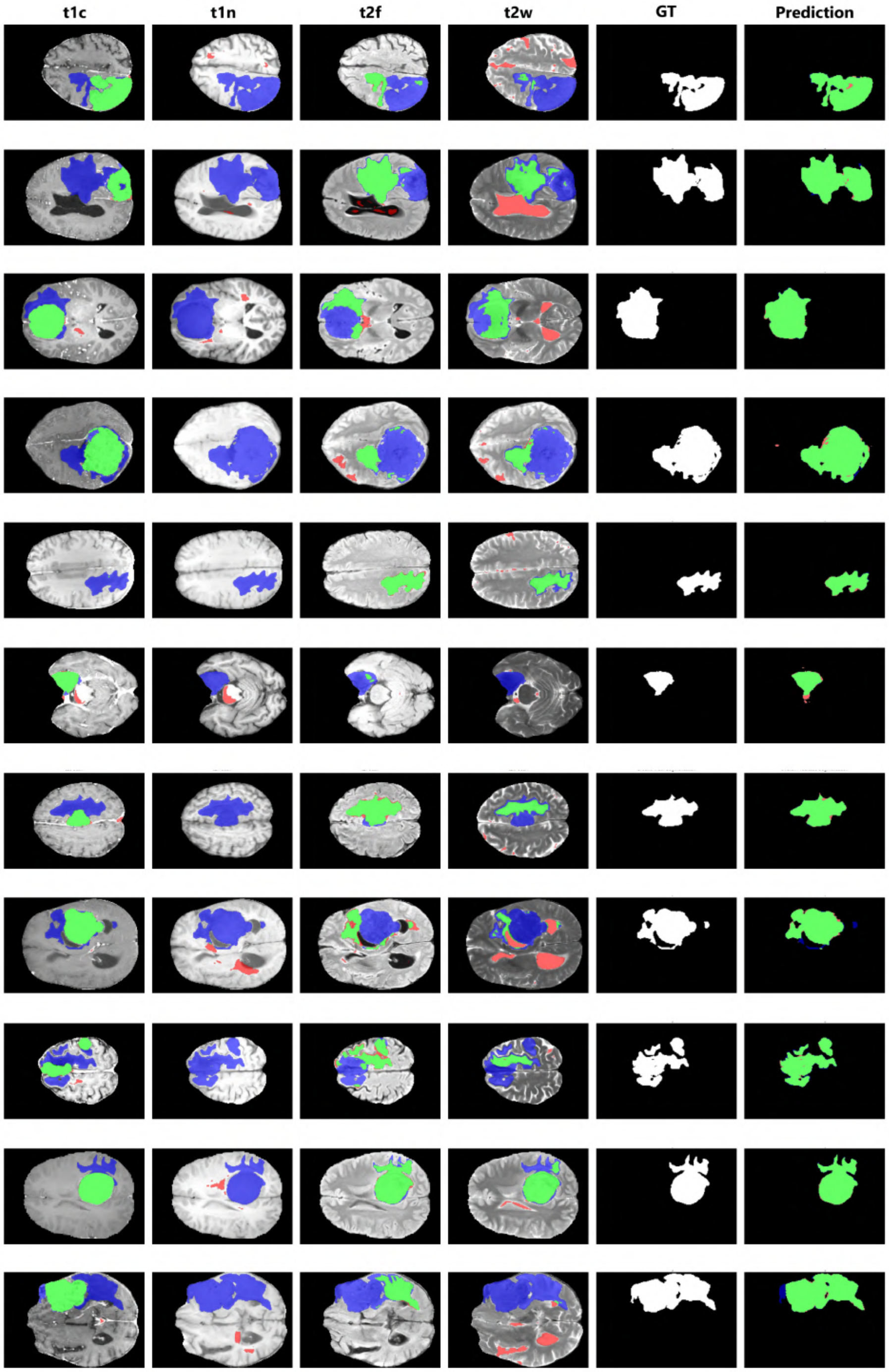


**Figure 14: Comparison of 3D U-Net (E100). The figure shows the per-scan predictions for each modality (t1c, t1n, t2f, t2w), the ground truth, and overall prediction, from left to right. This figure highlights the U-Net's precise edge delineation capabilities, which are notably better than those of Vanilla SAM. The model was trained on all scan types at once, making per-scan performance less relevant but still shown for clarity. The main takeaway is the U-Net's ability to achieve precise segmentation, even for complex tumours.**

## C U-SAM-COMBO AND U1-SAM2 VISUALS

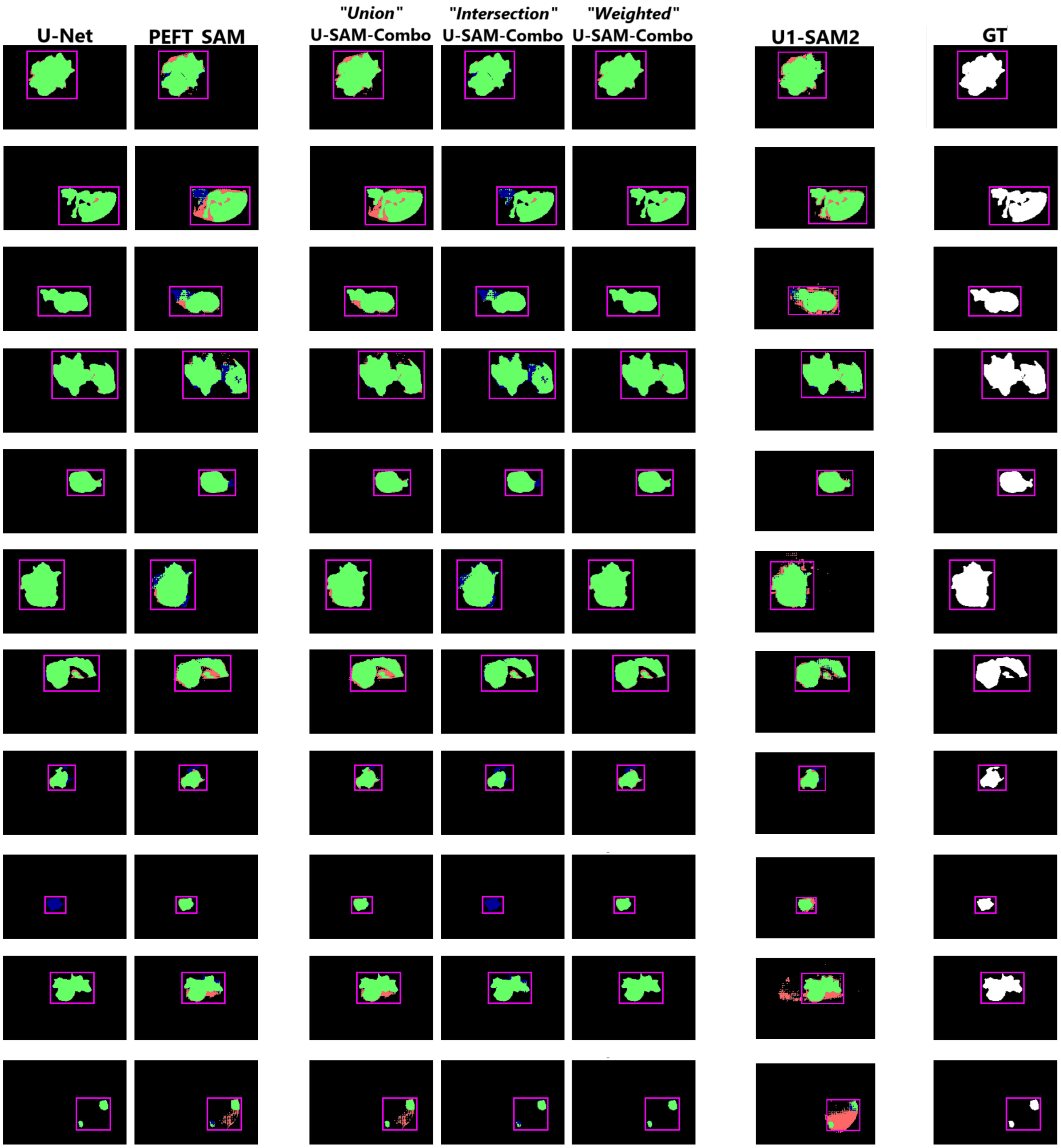


**Figure 15: Comparison of of U-SAM-Combo and U1-SAM2 models. Each row displays overall predictions of 3D U-Net (E100), PEFT-SAM (E125), U-SAM-Combo (Union, Intersection, Weighted), and U1-SAM2, followed by the ground truth, from left to right. This figure illustrates how combining different methods can significantly impact results. The Union and Weighted U-SAM-Combo methods show the most visually pleasing results, whereas U1-SAM2 exhibits noticeable spurious pixels. The subjective assessment demonstrates how combining methods can affect performance differently depending on the use case.**