Highlights

GBT-SAM: A Parameter-Efficient Depth-Aware Model for Generalizable Brain tumour Segmentation on mp-MRI

Cecilia Diana-Albelda[a,*], Roberto Alcover-Couso[a,b], Álvaro García-Martín[a], Jesus Bescos[a], Marcos Escudero-Viñolo[a]

- State-of-the-art results are achieved while significantly reducing computational costs through an efficient training strategy with fewer than 6.5M trainable parameters. This is accomplished by adding LoRA blocks and a slice selection approach that processes less than 2.6% of the slices.
- Utilization of all available modalities of multi-parametric MRI (mp-MRI) to enhance higher performance and enable generalization across different brain tumour domains.
- Improved volumetric segmentation is obtained by introducing a Depth-Condition block, which captures inter-slice correlations and enriches feature representations.

GBT-SAM: A Parameter-Efficient Depth-Aware Model for Generalizable Brain tumour Segmentation on mp-MRI

Cecilia Diana-Albelda[a,*], Roberto Alcover-Couso[a,b], Álvaro García-Martín[a], Jesus Bescos[a], Marcos Escudero-Viñolo[a]

[a]Universidad Autónoma de Madrid, Electronics and Communications Technology Dept., Madrid, 28049, Spain
[b]Amazon, Madrid, Spain, 28046, This work was done prior to joining Amazon.

Abstract

Gliomas are brain tumours that stand out for their highly lethal and aggressive nature, which demands a precise approach in their diagnosis. Medical image segmentation plays a crucial role in the evaluation and follow-up of these tumours, allowing specialists to analyse their morphology. However, existing methods for automatic glioma segmentation often lack generalization capability across other brain tumour domains, require extensive computational resources, or fail to fully utilize the multi-parametric MRI (mp-MRI) data used to delineate them. In this work, we introduce GBT-SAM, a novel Generalizable Brain Tumour (GBT) framework that extends the Segment Anything Model (SAM) to brain tumour segmentation tasks. Our method employs a two-step training protocol: first, fine-tuning the patch embedding layer to process the entire mp-MRI modalities, and second, incorporating parameter-efficient LoRA blocks and a Depth-Condition block into the Vision Transformer (ViT) to capture inter-slice correlations. GBT-SAM achieves state-of-the-art performance on the Adult Glioma dataset (Dice Score of 93.54) while demonstrating robust generalization across Meningioma, Pediatric Glioma, and Sub-Saharan Glioma datasets. Furthermore, GBT-SAM uses less than 6.5M trainable parameters, thus offering an efficient solution for brain tumour segmentation.
Our code and models are available at https://github.com/vpulab/med-sam-brain .

Keywords: Glioma, Brain tumour Segmentation, mp-MRI, SAM, LoRA, Parameter-efficient training.

1. Introduction

Gliomas, a type of primary brain tumour, rank among the most lethal forms of cancer, exhibiting notably low survival rates compared to other cancer types [1]. Accurate segmentation of these tumours is critical for diagnosis, enabling precise identification and delineation of the affected region. Currently, this task is performed manually by radiologists, which is both time-consuming and labour-intensive, ultimately increasing workload and delaying detection. Therefore, an automated mechanism to assist healthcare professionals in segmentation is essential for streamlining the process [2].

In the domain of automated tumour segmentation, Large Vision Models (LVMs) have shown remarkable progress over the past decade [3, 4], offering robust segmentation solutions in various contexts, including both medical [5, 6, 7] and non-medical applications [8, 9]. Among these, the Segment Anything Model (SAM) [10] stands out due to its strong generalization capabilities and its interactive segmentation approach, which allows experts to leverage it as an assistance tool accelerating their workflow without replacing their professional judgment. However, we identify three common limitations in current state-of-the-art segmentation systems:

First, most segmentation models are designed to process three-channel colour images [11, 12, 13]. This constraint is specially problematic for glioma imaging, with its most standardized imaging category being Multi-parametric Magnetic Resonance Imaging (mp-MRI), which comprises four distinct modalities [2, 14, 15, 16]: T1, T2, gadolinium contrast-enhanced T1 (T1c), and T2-FLAIR (Fluid-Attenuated Inversion Recovery). Each modality provides complementary information about the structure and characteristics of the tumours, making it crucial to utilize all of them for an accurate delineation [17]. To circumvent this issue, researchers often triplicate a single modality across three channels to satisfy model input requirements. However, this approach severely restricts the capabilities of the model, as no single modality can capture the full diversity of tumour presentations [18] (see Figure 1).

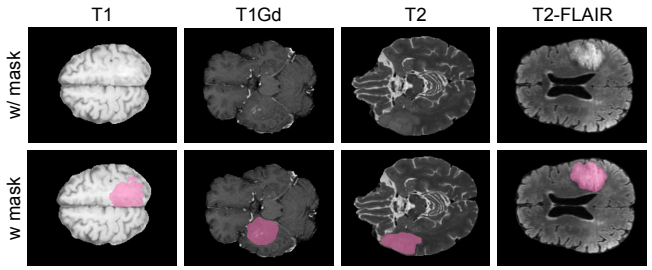

Figure 1: **MRI modalities.** Examples of tumour visualization in which each column represents a different MRI modality. In these cases, the full tumour extension is not visible in any single modality, highlighting the necessity of leveraging all of them to accurately predict and segment the tumour region. The top row shows the raw MRI images, while the bottom one includes the ground-truth tumour region.

Second, while MRI captures a 3D brain volume via a set of

slices, LVMs process each slice independently as an isolated image [10]. This approach may be sufficient for some domains, but it ignores the critical inter-slice correlations inherent in volumetric tumour images. To address this shortcoming, certain methods integrate attention mechanisms specifically designed to model these relationships, albeit at the cost of increased computational complexity [6, 18].

Third, LVMs generally originate from non-medical domains and therefore lack the specialized knowledge required for medical imaging tasks [4, 18, 19]. Consequently, these models must undergo domain adaptation procedures tailored to medical data. Moreover, they process all slices from the volumetric image simultaneously in the same batch during training without taking into account inter-slice correlations, leading them to positional biases. Overall, current mp-MRI training methods remain computationally expensive and are yet to be fully optimized.

To overcome the above mentioned limitations, we propose using Low-Rank Adapters and incorporating all four modalities of the complete mp-MRI to promote a generalizable model capable of segmenting brain tumours across diverse domains. Our proposed framework, Generalizable Brain Tumour SAM (GBT-SAM), includes the following contributions:

1. **Multi-modal Adaptation for mp-MRI**: We enable LVMs to simultaneously process all four mp-MRI modalities (T1, T2, T1c, and T2-FLAIR), thus providing the model with the same rich, complementary information that practitioners rely on. Rather than selecting a single modality and replicating it across channels, we modify the patch embedding of the model to accommodate a genuine 4-channel input. We further implement a two-stage training protocol to learn from the unique characteristics of each modality.
2. **Depth-Conditioned Correlation Modelling**: To address the challenge of inter-slice correlation, we introduce a novel Depth-Condition block. This component, strategically integrated at different levels of the model architecture, refines features by conditioning them on neighbouring slices. The result is an efficient mean of capturing volumetric context without high computational overhead.
3. **Parameter-Efficient Domain Adaptation**: Recognizing that current models often lack specialized medical domain knowledge, we implement a parameter-efficient approach to tune LVMs for the medical domain. Our technique streamlines the adaptation process by learning medical-specific features with reduced computational burden.

Building on these contributions, our approach requires up to 14 times fewer computational resources than the top-performing glioma segmentation method [20], while still achieving comparable performance, as illustrated in Figure 2. Additionally, to the best of our knowledge, this is the first method evaluated across four distinct brain tumour domains (Adult Glioma, Meningioma, Pediatric Glioma, and Sub-Saharan Gliomas), underscoring its broad applicability and adaptability to diverse clinical scenarios.

The content of this paper is organized as follows: Section 2 provides a background of gliomas biological fundamentals.

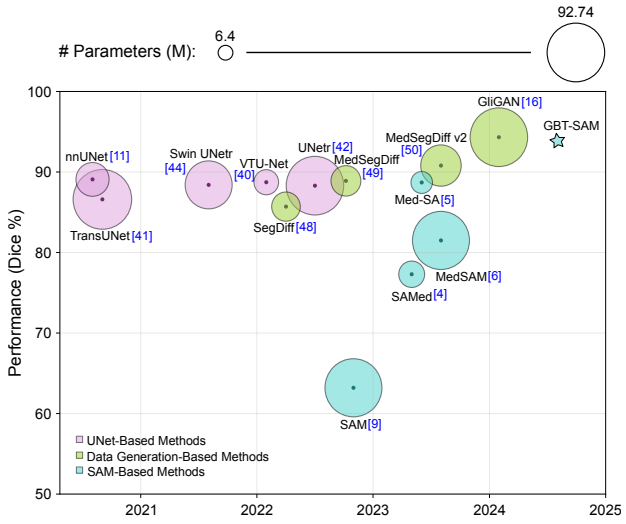

Figure 2: **Comparison of brain tumour segmentation methods based on performance (Dice Score), publication year, and model size (number of trainable parameters in millions)**. Pink, green and blue bubbles represent UNet-based, Generative, and SAM-based methods respectively. Our method, marked with a star, almost reaches the highest obtained Dice Score while training the smallest number of parameters, highlighting its efficiency and effectiveness compared to state-of-the-art approaches.

Section 3 reviews the current state of the art in tumour segmentation and LVM adaptation. Section 4 details our proposed architecture, GBT-SAM, including the multi-modal embedding strategy and Depth-Condition block. Section 5 presents our experimental setup, followed by quantitative and qualitative results. Also, we discuss the implications of our findings and outline potential areas of future research. Finally, we draw conclusions in Section 6.

2. Biological Fundamentals

In this section, we provide an overview of the biological and clinical background relevant to our study. First, we describe gliomas, focusing on their most aggressive variant, glioblastomas, to establish the biological context of our work. Next, we dive into domain variability, detailing the four brain tumour types used in this study: Adult Glioma, Meningioma, Pediatric Glioma, and Sub-Saharan Glioma, highlighting their unique characteristics. Following this, we explain the image acquisition process, providing insights into the specific features of the multi-parametric MRI (mp-MRI) modalities employed in this research. Lastly, we outline the annotation protocol to emphasize how medical experts rely on all MRI modalities at the same time for precise and comprehensive tumour delineation.

2.1. Glioma

Gliomas are primary tumours of the Central Nervous System (CNS) that originate from glial stem or progenitor cells [1]. They are classified into two main groups: 1) infiltrating or diffuse gliomas [21], which invade the surrounding brain tissue

extensively; and 2) circumscribed gliomas [22], which affect in a more localized manner.

Infiltrating gliomas are further classified into three main types according to their cellular origin [23]: astrocytomas [24], oligodendrogliomas [25] and oligoastrocytomas [26], each related to different types of glial cells. The World Health Organization assigns malignancy grades ranging from I (lower) to IV (higher) to these tumours, with glioblastoma (GB) being the grade IV infiltrating astrocytoma. GB is the most common and aggressive tumour among primary CNS malignancies in adults [27], characterized by extreme heterogeneity in appearance, shape and histology. Therefore, in this work we focus on taking GB cases for the Adult Glioma domain representation.

Despite advancements in the molecular understanding of gliomas, treatment remains reliant on surgery, radiotherapy, and chemotherapy, whose effectiveness depends heavily on the stage at which the disease is detected [28, 29, 30]. This underscores the importance of early detection, highlighting the need for AI-assisted tools that can expedite practitioners' workflows.

2.2. Brain Tumour Variability

Brain tumours encompass various categories, including Meningiomas and Gliomas. Meningiomas are predominantly benign tumours originating from arachnoid cap cells and account for 13–26% of all intracranial tumours. Anatomically, these tumours arise in different parts of the dura mater and can be found in areas such as the skull base and supratentorial regions. Gliomas, on the other hand, exhibit significant variability based on factors such as age group (Pediatric versus Adult) and the quality of the imaging technology used for acquisition.

In terms of age-related differences, Adult Gliomas exhibit rapid progression and resistance to conventional treatments. In contrast, Pediatric Gliomas display distinct biological characteristics [15], often demonstrating circumscribed growth and a favourable prognosis.

The imaging acquisition process also contributes to variability in glioma data. For example, Sub-Saharan Glioma imagery present particular properties due to the lack of high-quality health infrastructure and resources [16]. The low quality of these MRI images makes automated segmentation approaches difficult to implement without additional processing techniques.

Overall, the high variability existing within brain tumour imagery (see Figure 3) poses a considerable challenge to automatic brain tumour segmentation systems [31] as they should reliably be deployed across domains and age groups. To that end, our GBT-SAM is specifically tailored to preserve foundational knowledge of LVMs, yielding models robust to domain shifts. Moreover, we define a novel evaluation setup across four distinct brain tumour domains (Adult Glioma, Meningioma, Pediatric Glioma, and Sub-Saharan Gliomas) to ensure its reliability across the vast inter-variability of brain tumours.

2.3. Image Acquisition

The diagnosis of gliomas primarily depends on multiparametric MRI (mp-MRI) [17], which typically includes four modalities: T1, T2, T1 with contrast enhancement (T1c), and T2-FLAIR [2]. These modalities are acquired during a single MRI session per patient and offer complementary perspectives of the tissues, emphasizing different features through distinct molecular mechanisms. The MRI imaging process involves exposing the patient to a magnetic field generated by the MRI machine, along with intermittent radio-frequency pulses. These pulses primarily affect hydrogen ions (H+), which are abundant in water and fatty tissues, by changing their orientation and rotation frequency. Consequently, T1 MRI modality is obtained by measuring the time it takes for protons to return to their original longitudinal state after the interruption of a radio-frequency pulse, while T2 measures the time it takes for protons to de-synchronize before a new pulse [32]. Hence, T1 images emphasize fatty tissues, whereas T2 images accentuate fluids.

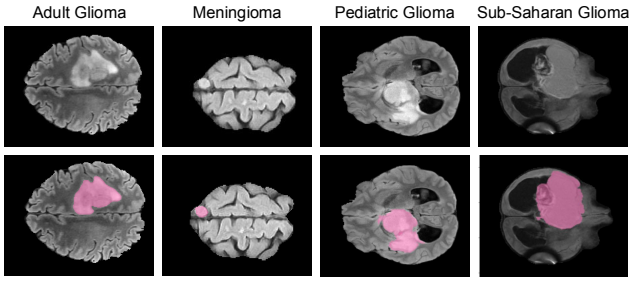

Figure 3: **Dataset.** Examples of each brain tumour domain used in the study: Adult Glioma, Meningioma, Pediatric Glioma, and Sub-Saharan Glioma. The first row displays the raw MRI images, while the second row highlights the tumour regions. Each dataset exhibits unique characteristics and variations in tumour appearance, underscoring the importance of a robust and generalizable segmentation model.

The T2-FLAIR modality is derived by suppressing cerebrospinal fluid signals in T2 images, enabling clearer visualization of other fluids, such as hemodynamics, that might otherwise be obscured [33]. Contrast-enhanced T1 images (T1c), on the other hand, are obtained by administering gadolinium, a contrast agent that enhances signals in areas with altered vascular permeability. This technique improves the detection of regions with increased vascularization, a hallmark of high-grade tumours such as glioblastomas (GB) [34].

Overall, each modality provides a unique perspective on brain tissues, making the analysis of all modalities essential for practitioners to make well-informed decisions.

2.4. Annotation Protocol

The brain tumour annotation protocol is conducted by radiologists, who simultaneously review all four mp-MRI modalities [2] following the detailed steps outlined below:

1. **Edema delineation**: Edema is an accumulation of fluid in brain tissue indicating swelling and increased pressure. This structure is initially identified on T2 images, which intensify the liquid tissues, and then refined on T2-FLAIR images, which suppress the cerebrospinal fluid signal.
2. **Detection of the active tumour**: The hyper-intensified areas in T1c are observed to identify the nucleus of the tu-

mour (core) and the active areas in expansion, which are usually the most vascularized and aggressive.

3. **Delimitation of necrotic areas**: Necrotic regions, which is the non-expanding part of the tumour core, are delimited by comparing the hypo-intense pixels in T1c with respect to T1 images.

This protocol ensures accurate segmentation of various tumour structures, offering essential information for diagnosis and treatment. We detail the annotation process to underscore a key limitation of current automatic segmentation systems: their reliance on a single mp-MRI modality for decision-making. In this paper, we propose a novel efficient method that integrates all four mp-MRI modalities, enhancing model performance.

3. Related Work

The most widely used methods for automatic medical imaging segmentation, including brain tumours, focus on three main approaches: UNet-based methods and their variants, data generation techniques i.e. diffusion models and generative adversarial networks (GANs), and foundational models such as SAM-based, which stand out for their interactive segmentation capabilities and greater generalization.

3.1. UNet-Based Methods

UNet-based models play a key role in the advancement of medical image segmentation, with ongoing innovations to address evolving challenges. The original UNet [35] architecture features a contracting path (encoder) to capture context and a symmetric expanding path (decoder) for accurate localization, enhanced by data augmentation to improve generalization. This design is extended in 3D UNet [36], replacing all 2D operations with 3D counterparts to better segment entire medical image volumes. Recognizing the unique characteristics of medical datasets, nnUNet [12] introduces a self-configurable approach that automatically adjusts the architecture, preprocessing, training, and post-processing for each task. This framework supports 2D and 3D UNet configurations as well as a cascade model, which refines a preliminary segmentation using a second UNet for high-resolution output. Similarly, DResUNet [37] incorporates deep residual networks to mitigate the vanishing gradient problem, combining low- and high-level features to accelerate training and enhance segmentation accuracy.

Attention mechanisms have further enhanced UNet-based designs. Attention UNet [38] incorporates Attention Gate (AG) modules that automatically focus on relevant structures, improving multi-class segmentation performance. VTUNet [39] extends this concept to 3D medical images, embedding Transformer blocks into the encoder and decoder to capture both local and global features. TransUNet [40] introduces transformers as context encoders for 2D images, paired with a UNet-like decoder for spatial information retrieval. This approach is further advanced in UNETR [41], which employs a transformer-based encoder linked to a decoder via skip-connections, enhancing global and multi-scale feature representation for 3D segmentation tasks. In parallel, the Swin Transformer [42] introduces a hierarchical shifted-window self-attention mechanism, which improves computational efficiency by focusing attention on local regions while enabling connections across them. Swin UNETR [43] builds on this concept for brain tumour semantic segmentation, utilizing hierarchical transformers to extract features across five resolutions, combined with a Fully-Connected Neural Network (FCNN) decoder through skip-connections. TransDoubleUNet [44] takes this further with a dual U-shaped architecture, integrating dual-scale Swin transformers and a hybrid Convolutional Neural Network (CNN)-transformer decoder to achieve highly accurate segmentation of multimodal brain tumour images.

Despite their successes, UNet-based enhanced with transformer architectures often struggle to generalize beyond the specific domains for which they are trained. Some studies attempt to overcome this by creating diverse, domain-spanning datasets to enhance model adaptability and robustness. However, these approaches require extensive data resources. To address these limitations, our work focuses on fine-tuning foundational vision models to ensure adaptability, even for unseen tumour domains.

3.2. Generative Methods

Data generation-based segmentation methods tackle the challenge of obtaining large, high-quality datasets by utilizing synthetic data or modifying existing datasets. Two prominent approaches in this domain are diffusion models and GANs, each offering distinct advancements and contributions.

Diffusion methods, such as De-noising Diffusion Probabilistic Models (DDPMs) [45], are generative models that use Markov chains to transform simple distributions, such as Gaussian noise, into complex data. The forward diffusion process progressively adds noise to the data, while the inverse process employs an optimized variational framework to reconstruct high-quality data [46]. SegDiff [47] extends diffusion models for image segmentation by integrating information from the original image with the evolving segmentation map. Using a UNet-based network, SegDiff iteratively refines the segmentation map through probabilistic diffusion processes. MedSegDiff [48] further adapts DDPMs for medical segmentation by introducing two key components: Dynamic Conditional Encoding (DCE), which adjusts conditions adaptively at each sampling step, and the Feature Frequency Parser (FF-Parser), which removes high-frequency noise. These enhancements improve regional focus and ensure the retention of critical features during the reverse diffusion process. Advancements like MedSegDiff-v2 [49] incorporate transformers into the diffusion model framework. Its Spectrum-Space Transformer (SS-Former) employs frequency-domain cross-attention to dynamically align noise and semantic features during each step, resulting in refined segmentation quality.

GANs have also demonstrated effectiveness in generating synthetic data for medical segmentation. GliGAN [20], for instance, employs an architecture with Swin UNETR as the generator and a simple CNN as the discriminator. This model creates synthetic tumours randomly placed in healthy brain regions and combines the synthetic data with real images, improving

prediction robustness. Other studies explore novel augmentation methods using multi-armed bandits for dynamic image deformations [50].

Overall, image generation-based models have made notable progress in recent years. However, they still face significant limitations, as the generated images often fall short of the realism compared to real images [4, 51, 52]. Additionally, their training and implementation demand substantial computational resources due to their complexity. To address these challenges, this paper focuses on adapting foundational models for brain tumour image segmentation, offering robust generalization across diverse domains with limited data.

3.3. SAM-Based Methods

The Segment Anything Model (SAM) [10] is a foundational model designed to deliver zero-shot segmentation performance across various visual domains. Its framework includes a prompt module that provides spatial cues, enabling the segmentation of specific objects within an image. This human-guided segmentation capability is particularly appealing for radiologists, as it allows them to provide initial information about tumour localization that can be further refined, offering a promising pathway for SAM to be integrated as a clinical tool to expedite tumour segmentation. However, as a generalist model, its performance in medical imaging is suboptimal due to the significant differences between natural and medical imaging domains. Challenges such as low contrast, poorly defined boundaries and small lesion regions limit its effectiveness in this context [11, 13]. To overcome these limitations, several strategies have been developed to adapt SAM for medical tasks.

To address the domain gap, methods such as SAM-Med2D [53] and MedSAM [7] perform full model fine-tuning using large medical datasets. However, these techniques are computationally expensive and time-consuming, limiting their scalability. In order to reduce computational costs, SAM-U [54] proposes a no-training method that contrasts predictions through multiple prompts to discard non-accurate predictions, however, as the model is not adapted to the target scenario, its performance is limited. Therefore, efficient training protocols arise as a plausible solution for SAM adaptation: SAMed [5] introduces a more efficient strategy based on Low-Rank adaptation (LoRA), fine-tuning the image encoder, prompt encoder and mask decoder, achieving effective adaptations with fewer resources. In this direction, Med-SA [6] proposes the Medical SAM Adapter, which incorporates domain-specific medical knowledge through light-weight adaptations.

Another limitation of SAM, designed for natural RGB images, is that it does not account for the volumetric nature of brain tumour imagery. To address this, Med-SA [6] introduces the Space-Depth Transpose (SD-Trans) technique, which extends SAM to 3D medical imaging by capturing correlations between volumetric slices. However, this approach comes with significant computational costs, as it relies on processing the entire volumetric image. Alternatively, Medical SAM 2 [55] explores the use of SAM 2 [56], a model designed for videos, which is an alternative source of volumetric data. Medical SAM 2 not only applies segmentation to 3D medical images but also introduces an innovative single-prompt segmentation capability. After an initial prompt, the model can automatically identify and segment similar objects in subsequent slices, optimizing the process. While volumetric images can be treated as analogous to video data, SAM 2 is not fully optimized for such sequences and incurs higher computational costs.

Based on the aforementioned limitations, this paper focuses on two key advancements: (1) the efficient adaptation of SAM using Parameter Efficient Fine-Tuning (PEFT) tailored to brain tumour imagery, and (2) the introduction of a novel Depth-Condition module that enhances segmentation by integrating information from adjacent upper and lower slices into the decision-making process. Our approach is highly efficient compared to its predecessors, requiring less than 2.6% of the available slices to effectively model correlations between them.

4. Method

In this section, we first formulate the task of brain tumour segmentation as a volumetric image segmentation framework. Second, we describe the general SAM pipeline which is our building block. Third, we introduce the proposed modifications to the patch embedding to combine information of all mp-MRI modalities, followed by a tailored training setup to effectively adapt the modified layers of SAM. Fourth, we introduce the employed parameter-efficient fine-tuning based on LoRA blocks. Finally, we describe the proposed Depth-Condition block to capture inter-slice correlations. All together, these contributions describe the proposed GBT-SAM pipeline for mp-MRI brain tumour image segmentation, illustrated in Figure 4.

4.1. Volumetric Image Segmentation

The volumetric brain tumour segmentation task in this study aims to assign a binary label to each voxel in mp-MRI, indicating whether it belongs to the tumoural region or not [57, 58]. Consider a mp-MRI volume $X \in \mathbb{R}^{H\times W\times D\times M}$, where H and W represent the height and width of each slice, D is the number of slices, and M represents the different MRI modalities T1, T2, T1c, and T2-FLAIR.

The goal is to create a model, parametrized by θ, that assigns a tumour probability to each voxel. We define:

$$f_\theta : \mathbb{R}^{H\times W\times D\times M} \to [0,1]^{H\times W\times D\times M}, \tag{1}$$

such that

$$Y_{\text{pred}} = f_\theta(X), \tag{2}$$

where $Y_{\text{pred}} \in [0,1]^{H\times W\times D\times M}$ is the predicted probability of the whole volumes and $y_{\text{pred}}(h,w,d,m)$ denotes the probability of a specific voxel being part of the tumour. On the other hand, $Y \in \{0,1\}^{H\times W\times D\times M}$ is the ground-truth segmentation, where $y(h,w,d,m) = 1$ indicates that the voxel belongs to the tumour region. Hence, the training problem consists in adjusting θ so that Y_{pred} closely matches Y.

To quantify the discrepancy between predictions and ground-truth, we employ the Binary Cross-Entropy (BCE) loss [59],

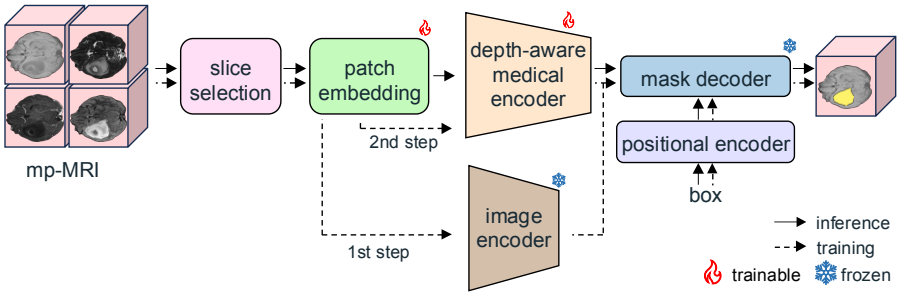

Figure 4: **GBT-SAM pipeline**. In a first training step, we perform slice selection to reduce computational costs while enhancing generalization capability. Moreover, the patch embedding layer is trained, while the rest of the modules remain frozen: image encoder, responsible for extracting features from the input slices; positional encoder, combining features with the bounding box information; and mask decoder, producing the predicted segmentation. In a second training step, the patch embedding layer is further trained alongside additional trainable components introduced in a modified version of the image encoder (depth-aware medical encoder): LoRA blocks and a Depth-Condition module.

which is the original SAM loss [60], hence is preferred for its fine-tuning. The BCE loss can be defined as:

$$\mathcal{L}_{\text{BCE}}(Y, Y_{\text{pred}}) = -[\, Y \log(Y_{\text{pred}}) + (1 - Y) \log(1 - Y_{\text{pred}})]. \quad (3)$$

In this manner, the volumetric tumour segmentation problem is framed as a minimization of the BCE loss function, aiming to find a model that accurately assigns tumour probabilities to each voxel of the mp-MRI volume.

4.2. SAM Pipeline

GBT-SAM pipeline builds on top of SAM to perform volumetric brain tumour segmentation taking as input mp-MRI volumes $X \in \mathbb{R}^{H \times W \times D \times M}$, as defined above. Given the size and complexity of the input data, a slice sampling strategy is implemented to reduce computational costs by selecting a random slice X_D each iteration and three additional ones at a fixed distance δ from each other, as defined in Equation 4, where $X_D \in \mathbb{R}^{H \times W \times M}$ and $\delta \in \{1, 4, 10\}$. This approach not only reduces computational requirements by focusing on a limited number of slices per iteration but also aims to enhance the generalization capabilities of the model by preventing it from relying on complete volume information at once.

$$X = \{X_{D-\delta}, X_D, X_{D+\delta}, X_{D+2\delta}\}, \quad (4)$$

These selected slices are passed through the patch embedding block, which partitions them into smaller patches [61], transforming each into a fixed-dimensional vector representation. The embedded patches are then fed into the ViT blocks of the image encoder, which extracts hierarchical features that capture both local and global contextual information [62]. These features are enriched with a Depth-Condition block designed to take into account the correlation between slices, hence providing a robust representation of the volumes.

The extracted features are subsequently combined with a bounding box prompt encoding, which supplies spatial guidance regarding the location of the tumour, according to interactive segmentation guidelines of SAM [63]. The combination of contextual and spatial information is processed by the mask decoder to produce a segmentation mask that delineates the final predicted tumour region.

This pipeline is designed to maximize efficiency by incorporating slice selection and leveraging the architecture to handle volumetric mp-MRI data, promoting accurate tumour segmentation while minimizing computational overhead.

4.3. Combining Information of All Modalities

SAM is primarily designed to process RGB images, with its patch embedding layer configured to accept 3-channel inputs. Given that our data consists of 4D mp-MRI, a direct adaptation is necessary to fully utilize this multi-modal volumes.

Existing approaches typically address this limitation by replicating one of the MRI modalities across all three channels [12], therefore underutilizing the given data and discarding valuable information [64]. GBT-SAM modifies the patch embedding input layer to accept 4-channel entrances instead of 3 [18]. Formally, let the patch embedding be represented as a function:

$$P : \mathbb{R}^{H \times W \times M} \rightarrow \mathbb{R}^{N \times d}, \quad (5)$$

where N is the number of patches generated from the input image, M is the number of channels, and d is the dimensionality of the feature vector for each patch. In the modified architecture, the patch embedding layer processes all four channels simultaneously, with each channel corresponding to one of the MRI modalities. Thus, the input to P is $X_D \in \mathbb{R}^{H \times W \times 4}$, where X_D is a slice selected from the volume.

This adaptation ensures that the full multi-modal information from mp-MRI is leveraged, mirroring the annotation process used by medical professionals, where each MRI modality provides complementary insights into the tumour structure.

4.4. Training Setup

The training of the GBT-SAM follows a two-step protocol to achieve efficient adaptation of SAM to the task of brain tumour segmentation while ensuring optimal utilization of the modified architecture. As the initial layer of the patch embedding is modified and initialized randomly so that it fits M channel

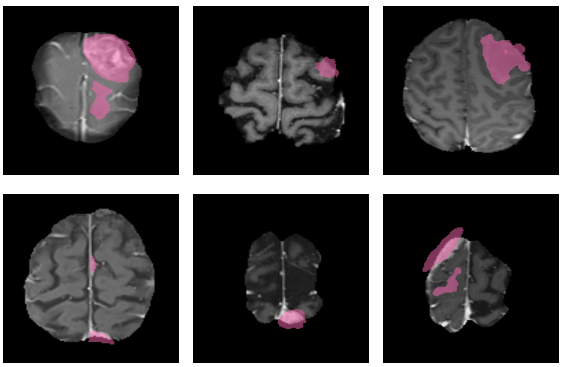

Figure 5: **Ground-truth incongruity due to slice correlation**. Examples of MRI slices where some pixels are annotated as tumour regions (highlighted in pink) despite the tumour or brain structure is not visible in the pixel. This is due to the fact that doctors may analyse contiguous slices simultaneously, leveraging the volumetric context, hence highlighting the importance of enriching image features by incorporating inter-slice correlations.

inputs, we propose a sequential training to avoid compensatory learning of encoder layers.

In the first step, the entire architecture is frozen except for the patch embedding layer, which processes the full mp-MRI input, capturing information from all four MRI modalities. By training only the patch embedding layer, the model leverages the generalization capabilities of the pre-trained SAM image encoder while adapting to the specific characteristics of the input data, thus ensuring that the patch embedding layer is properly optimized to handle the multi-modal nature of the data before introducing additional trainable parameters.

In the second step, we employ a parameter-efficient fine-tuning strategy [65] in which the patch embedding layer remains trainable and LoRA blocks [66] are introduced within the ViT modules of the image encoder (see Section 4.5). These LoRA blocks allow the architecture to acquire medical domain-specific knowledge without modifying the entire SAM encoder, thereby significantly reducing the computational cost [67, 5]. Additionally, a Depth-Condition block is integrated to capture inter-slice correlations [68] (see Section 4.6).

This two-step training process is designed to ensure a gradual and efficient adaptation of the pipeline. Training the patch embedding layer in isolation during the first step prevents the LoRA blocks from learning to compensate for errors in the patch embedding layer. Consequently, the second step focuses on fine-tuning the system as a whole, achieving a better combination of the image encoder and the modified patch embedding.

4.5. Parameter-Efficient Fine-Tuning

To efficiently adapt SAM to the medical domain, we employ a Parameter-Efficient Fine-Tuning (PEFT) approach [65] based on Low-Rank Adaptation (LoRA) [66]. Consider the set of parameters θ that define the segmentation model f_θ. A direct fine-tuning of θ over all dimensions would be computationally expensive and could lead to suboptimal performance without large-scale medical data [69]. Instead, we decompose the parameter update $\Delta\theta_l$ for a given layer l into two low-rank matrices. Being $\theta_l \in \mathbb{R}^{N\times L}$ the original SAM weights for layer l, we define the updated weights as:

$$\Theta_l = \theta_l + \Delta\theta_l. \tag{6}$$

The objective is to constrain Θ_l to lie in a low-dimensional subspace. Specifically, we factorize:

$$\Delta\theta_l = AB, \tag{7}$$

where $A \in \mathbb{R}^{N\times r}$ and $B \in \mathbb{R}^{r\times L}$, with r controlling the rank of the matrices A and B.

During training, we keep θ_l fixed and only optimize the entries of A and B. Following [66], we initialize $A \sim \mathcal{N}(0, \sigma)$ and $B = \mathbf{0}$. The layer l then produces its output as:

$$\Theta_l(x) = \theta_l x + ABx. \tag{8}$$

This factorization restricts the update to a subspace, hence significantly reducing the number of trainable parameters acting as a form of regularization. By doing so, we preserve the foundational structure of SAM while enabling it to specialize to the task of volumetric brain tumour segmentation with minimal additional computational cost.

4.6. Depth-Condition Block

When dealing with volumetric medical data, it is crucial to incorporate mechanisms that account for the correlations between features across slices [6]. Unlike natural images, where each image is typically processed independently, medical volumes such as mp-MRI provide complementary information distributed throughout the volumetric structure [32]. As shown in Figure 5, there are cases in which tumour regions are annotated in pixels where the tumour is not directly visible, or even where the brain itself is absent, specially for the first or last slices of the volumes. This phenomenon arises because medical experts often annotate tumour regions by simultaneously observing a few slices. Ignoring these correlations by processing slices as independent images can lead to a loss of critical contextual information, thus it is very important that image features capture inter-slice correlations, hence ensuring that the model identifies the dependencies inherent in volumetric data [70, 71].

To address this, we introduce a Depth-Condition module in each ViT block of the image encoder. This module, represented in Figure 6, operates directly on the volumetric representation $X \in \mathbb{R}^{H\times W\times D}$ by modelling interactions between slices.

The block begins by unfolding the volumetric data along the slice dimension D, effectively reshaping the volume into a format that facilitates depth-wise operations while retaining spatial integrity. A layer normalization step is then applied to standardize the input features [72], and then the data is passed through a simple Multi-Layer Perceptron (MLP) designed to learn dependencies and interactions across slices.

After MLP, the features are folded back into their original volumetric structure, enriched with depth-aware representations that reflect the interactions learned. These Depth-Conditioned features are seamlessly integrated into the ViT

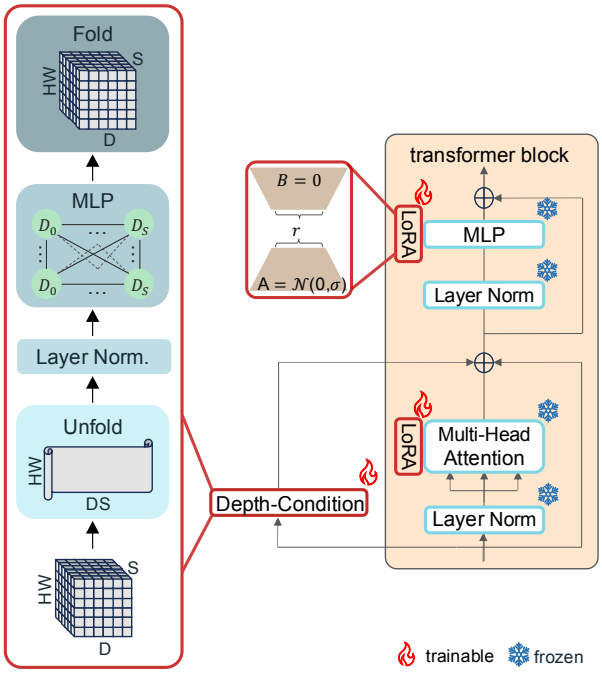

Figure 6: **Depth-Condition Block.** The Depth-Condition block consists of three main stages: unfolding the volumetric data along the slice dimension with layer normalization, processing it to extract depth-specific features, and folding the data back into the volumetric structure. This Depth-Conditioned output is then integrated into the ViT block, where it complements the features processed by the LoRA modules, enabling the model to leverage both spatial and volumetric information.

blocks in conjunction with LoRA modules, where they complement the spatial information captured by other components of the model. This integration allows the transformer to leverage both spatial and inter-slice dependencies for improved segmentation accuracy.

5. Experiments & Results

In this section, we describe the experimental setup for the study, including dataset characteristics of the different domains, evaluation metrics and computational resources required to perform both training and evaluation phases. Additionally, an extensive evaluation of GBT-SAM has been performed through a detailed ablation study to analyse the impact of significant hyperparameters on the performance of the model. Moreover, we compare our method against current state-of-the-art approaches for brain tumour segmentation, demonstrating its good performance and efficiency. Finally, we provide visual results to illustrate the performance of our model, showcasing predictions alongside ground-truth annotations for a qualitative assessment.

5.1. Experimental Setup

5.1.1. Datasets

The datasets used in this study have been extracted from the public RSNA-ASNR-MICCAI BraTS 2023 challenge [2], which aims to promote research on automatic methods for the diagnosis of gliomas. Our study extends to four datasets corresponding to different brain tumour domains $D = \{D_1, D_2, D_3, D_4\}$, where:

- D_1 (Adult Glioma) = $\{n_0, n_1, ..., n_{1251}\}$ [2]
- D_2 (Meningioma) = $\{n_0, n_1, ..., n_{1000}\}$ [14]
- D_3 (Pediatric Glioma) = $\{n_0, n_1, ..., n_{99}\}$ [15]
- D_4 (Sub-Saharan Glioma) = $\{n_0, n_1, ..., n_{60}\}$ [16],

being each n_j a patient's full mp-MRI scan. As state-of-the-art methods primarily report their performance on the D_1 dataset, for a fair comparison, we limit our training to D_1. However, we evaluate the generalization capability of GBT-SAM by testing the trained model on all four datasets (D_1, D_2, D_3, D_4), as well as by including the mean performance on unseen datasets (DS_{234}), defined in Equation 10. This strategy allows us to assess how effectively the model adapts to diverse brain tumour domains, highlighting its robustness and versatility.

5.1.2. Evaluation Metrics

For all experiments presented in this section, the Dice Score (DS) [73] is employed as the evaluation metric, defined as:

$$DS_i = \frac{2 \cdot |Y_i \cap Y_{\text{pred,i}}|}{|Y_i| + |Y_{\text{pred,i}}|}, \tag{9}$$

where Y represents the ground-truth segmentation, Y_{pred} denotes the predicted segmentation and the subindex i indicates the specific dataset of D. This metric is the most used in brain tumour segmentation and it ensures consistent and reliable performance comparison [74].

Furthermore, the mean of the DS metric for all unseen datasets is calculated as follows in order to evaluate the generalization capability of the model:

$$DS_{234} = \frac{DS_2 + DS_3 + DS_4}{3} \tag{10}$$

5.1.3. Computational Resources

Each training session was executed on a single 48GB NVIDIA A40 GPU, utilizing a batch size of 4 and Adam optimizer [75]. We employed an 80-20 split of the dataset to validate the learning outcomes for each case.

5.2. Ablation Study

An extensive ablation study has been conducted to analyse the impact of various hyperparameters on the performance of GBT-SAM. By doing this, we aim to maximize the segmentation accuracy and generalization capabilities.

All experiments in the ablation study are evaluated based on the DS for our four available brain tumour domains: D_1 [2], D_2 [14], D_3 [15] and D_4 [16]. This evaluation allows us to assess the adaptability of the proposed method across diverse clinical scenarios and data distributions.

First of all, we explore different values for the rank (r) of the matrices introduced by the LoRA blocks [66], as this component is specifically designed to provide medical domain knowledge to the network [76]. The results of this experiment are summarized in Table 1.

	$DS_1 \uparrow$	$DS_2 \uparrow$	$DS_3 \uparrow$	$DS_4 \uparrow$	$DS_{234} \uparrow$
r=4	82.52	80.62	79.25	78.18	79.35
r=8	84.13	78.79	79.44	75.43	77.89
r=12	85.09	80.88	**83.51**	**79.00**	**81.13**
r=16	**86.11**	**80.91**	80.50	75.95	79.12

Table 1: **LoRA rank.** Impact of rank r in the LoRA blocks on DS across Adult Glioma (DS_1), Meningioma (DS_2), Pediatric Glioma (DS_3), and Sub-Saharan Glioma (DS_4). Increasing r improves DS for the training set D_1, with $r = 16$ achieving the best performance. However, $r = 12$ gets better generalization capabilities (DS_{234}).

From the table, we observe that increasing the rank r generally leads to improved performance for the training domain (D_1), with $r = 16$ achieving the highest DS on this dataset. However, when evaluating generalization capability across unseen brain tumour domains (DS_{234}), $r = 12$ emerges as the optimal choice, providing the best overall balance between segmentation accuracy and generalization. Moreover, the difference in performance on D_1 between $r = 16$ and $r = 12$ is not significant, which also justifies our decision to use $r = 12$ as the default rank in subsequent experiments to ensure robustness across diverse brain tumour domains.

Once the network has been equipped with medical-specific knowledge through LoRA blocks, we evaluate the impact of utilizing the complete mp-MRI volume compared to two alternative approaches: replicating a single modality, as commonly done in RGB-based systems, or combining only the three most representative MRI modalities, i.e. those yielding the highest DS among the four available ones.

MRI Mod.	$DS_1 \uparrow$	$DS_2 \uparrow$	$DS_3 \uparrow$	$DS_4 \uparrow$	$DS_{234} \uparrow$
T1	49.47	68.96	57.65	42.15	56.25
T1c	71.27	76.96	58.36	51.09	62.14
T2	81.86	73.80	73.14	64.18	70.37
T2-FLAIR	82.62	79.83	83.45	70.18	77.95
[T1c, T2, T2-FLAIR]	**86.84**	79.99	77.52	76.03	77.85
All mp-MRI	85.09	**80.88**	**83.51**	**79.00**	**81.13**

Table 2: **1 vs 3 vs 4 MRI Modalities.** Comparison of replicating individual modalities, combining the three most representative ones, and our proposed approach using all four modalities. Our method leveraging the entire mp-MRI obtains second best DS on Adult Glioma (DS_1), which is the training set, and the best generalization capability across all not seen brain tumour domains (DS_{234}): Meningioma, Pediatric Glioma, and Sub-Saharan Glioma.

From the results shown in Table 2, it is evident that relying on individual modalities significantly underperforms compared to using a combination of modalities or the full mp-MRI. While combining three representative modalities achieves competitive results on D_1 (Adult Glioma), it falls short in generalization to other domains, particularly on D_3 (Pediatric Glioma) and D_4 (Sub-Saharan Glioma). Conversely, leveraging all four MRI modalities demonstrates superior generalization capability DS_{234}, while maintaining strong performance on D_1.

These results highlight that the inclusion of all four mp-MRI modalities allows the network to capture a more comprehensive understanding of tumour morphology and characteristics across varying brain tumour domains, hence confirming the importance of fully utilizing the mp-MRI data.

The next step involves defining the slice selection block to optimize how the data is provided to the network. To achieve this, we evaluate two strategies: selecting 4 distinct random slices per iteration or selecting 1 random slice followed by 3 additional ones at a fixed distance δ (see Equation 4).

	$DS_1 \uparrow$	$DS_2 \uparrow$	$DS_3 \uparrow$	$DS_4 \uparrow$	$DS_{234} \uparrow$
Random	83.95	79.21	70.67	69.12	73.00
δ=10	83.86	78.62	78.57	75.60	77.60
δ=4	85.09	80.88	**83.51**	79.00	81.13
δ=1	**88.81**	**85.49**	83.46	**80.60**	**83.18**

Table 3: **Slice selection strategies.** We explore taking 4 random slices or 1 random slice combined with other 3 at a fixed distance ($\delta = \{10, 4, 1\}$). DS across Adult Glioma (D_1), Meningioma (D_2), Pediatric Glioma (D_3), and Sub-Saharan Glioma (D_4) shows that selecting 4 consecutive slices ($\delta = 1$) yields the best performance.

As shown in Table 3, the architecture achieves the best performance when using slices with $\delta = 1$, surpassing all other configurations. This highlights the importance of feeding the model sequential images, as consecutive slices provide complementary spatial information, thus enhancing the ability to capture inter-slice correlations. Consequently, the choice of $\delta = 1$ is adopted in subsequent experiments to achieve optimal segmentation performance.

	$DS_1 \uparrow$	$DS_2 \uparrow$	$DS_3 \uparrow$	$DS_4 \uparrow$	$DS_{234} \uparrow$
1p	88.11	85.49	83.46	80.60	83.18
BB-100-75	90.90	**91.85**	90.77	87.24	89.95
BB-75-75	**91.35**	91.31	**91.18**	**88.59**	**90.36**
BB-100-100	92.17	92.68	91.54	90.62	91.61

Table 4: **Prompts Comparison.** Results for 1 point prompt ($1p$) and bounding box prompts (BB) with varying overlap percentages during training and testing (BB-100-75 and BB-75-75). With BB-100-100 indicating the unrealistic upper DS bound across all domains (Adult Glioma, D_1; Meningioma, D_2; Pediatric Glioma, D_3; and Sub-Saharan Glioma, D_4), we select BB-75-75 as the most strong and practical option.

As with all state-of-the-art SAM-based methods, which rely on prompts to harness their interactive segmentation capabilities, in Table 4 we explore different types of prompts. Initially, we evaluate the use of a single point prompt ($1p$) to guide the segmentation process. Subsequently, we investigate bounding box prompts (BB) with varying overlap percentages with the tumour region. Specifically, we experiment with BB-100-100, where bounding boxes have 100% overlap with the tumour during both training and testing; BB-100-75, where it is maintained 100% overlap during training but reduced to 75% for testing; and BB-75-75, where 75% is used in both training and testing.

From the results, it is clear, as expected, that BB-100-100 achieves the highest performance in all datasets, indicating the upper performance bound for the method. However, this setting

is impractical and unrealistic for clinical use, as it assumes perfect annotations. Therefore, we select the next best-performing option, *BB*-75-75, which achieves strong and consistent results while aligning more closely with real-world scenarios.

Lastly, we evaluate different training protocols to determine the most effective approach for the network. The three strategies assessed are: (1) training solely the patch embedding layer, to investigate whether this alone provides sufficient medical domain information; (2) a one-step training protocol where the patch embedding layer and LoRA blocks are trained simultaneously; and (3) a two-step training protocol, where the patch embedding layer is trained first, followed by a second step in which it is retrained alongside LoRA blocks. Each of these protocols is further evaluated with and without the inclusion of our proposed Depth-Condition block, which enriches the model by capturing inter-slice correlations.

Training Process	Slice Corr.	$DS_1 \uparrow$	$DS_2 \uparrow$	$DS_3 \uparrow$	$DS_4 \uparrow$	$DS_{234} \uparrow$
Patch Embed.		89.70	88.59	90.74	87.70	89.01
1-step	w/o D.C	91.35	91.31	91.18	88.59	90.36
	D.C	92.82	91.33	92.10	90.20	91.21
2-step	w/o D.C	92.10	**91.88**	90.08	89.54	90.5
	D.C	**93.54**	91.87	**91.85**	**91.07**	**91.60**

Table 5: **Training Strategy.** D.C is related to the use our proposed Depth-Condition block, while w/o D.C represents not using it. The table evaluates: (1) training only the patch embedding layer, (2) one-step training of patch embedding and LoRA blocks together, and (3) two-step training process, first patch embedding and then patch embedding with LoRA blocks. The two-step process with D.C achieves the best DS (DS_1, DS_2, DS_3, D_4 and D_{234}) across all datasets (Adult Glioma, Meningioma, Pediatric Glioma and Sub-Saharan Glioma).

The results presented in Table 5 demonstrate that the two-step training protocol combined with the Depth-Condition block achieves the best performance across all brain tumour studied scenarios (D_1, D_2, D_3, D_4). This configuration maximizes segmentation performance by ensuring a robust initial adaptation to mp-MRI data during the first step and enhancing the feature representation during the second step. These findings highlight the importance of separating the training stages to prevent compensatory learning and the critical role of the Depth-Condition block in leveraging inter-slice dependencies for improved volumetric segmentation.

5.3. Comparison with State-of-the-Art

We compare GBT-SAM with current state-of-the-art approaches for brain tumour segmentation. All evaluated methods are assessed solely in terms of DS on the Adult Glioma dataset (DS_1), as this is the most extensively studied domain in the literature [37, 41]. Moreover, to the best of our knowledge, our method is the first to not only address Adult Glioma segmentation but also explore and compare its generalization capability across Meningioma, Pediatric Glioma, and Sub-Saharan Glioma. Consequently, existing state-of-the-art methods do not provide results or comparisons for the additional datasets included in this study. Hence, DS_1 would be the only shared benchmark for a fair comparison.

The results of the comparison between GBT-SAM and current state-of-the-art methods for brain tumour segmentation on the Adult Glioma dataset (D_1) are presented in Table 6. Among the promptable SAM-based approaches, GBT-SAM achieves the highest DS (93.54), significantly surpassing all other models in the same family. This demonstrates the efficacy of our modifications to SAM for adapting it to the specified domain.

Method	*Algorithm Family*	$DS_1 \uparrow$
3D UNet [36]		84.11
Att. UNet [38]		85.57
DResU-Net [37]		86.60
TransUNet [40]		86.60
UNetr [41]	*UNet*	88.30
Swin-UNetr [43]		88.40
nnUNet [12]		88.50
VTU-Net [39]		88.73
TransDoubleU-Net [44]		92.87
SegDiff [47]		85.7
MedSegDiff [48]	*Data Generation*	88.90
MedSegDiff v2 [49]		90.80
GliGAN [20]		**94.32**
SAM [10]		63.20
SAMed [5]		77.30
SAM-U [54]		81.0
MedSAM [7]	*SAM*	81.50
SAM-Med2D [53]		82.90
Med-SA [6]		88.70
Medical SAM-2 [55]		90.50
GBT-SAM		93.54

Table 6: **Comparison of GBT-SAM with state-of-the-art for brain tumour segmentation in terms of DS on the Adult Glioma dataset** (DS_1). Our method achieves the best performance among SAM-based approaches and ranks second overall, demonstrating its effectiveness and competitiveness against existing methods.

Compared to methods based on UNet architectures, GBT-SAM also outperforms all of the models, surpassing even very popular implementations like Swin-UNetr [43] (88.40) and nnUNet [12] (88.50), being only TransDoubleU-Net [44] (92.28) close in performance within this category. Additionally, GBT-SAM achieves the second top performance with respect to generative methods, with only GliGAN [20] beating it (94.32).

Overall, GBT-SAM ranks second across all tested methods, which validates the design of our proposed pipeline and its enhancements for handling brain tumour segmentation tasks.

To further analyse the efficiency of segmentation methods, we also provide a comparison of GBT-SAM with state-of-the-art approaches that utilize fewer than 50 million trainable parameters. The results, summarized in Table 7, provide insights into the trade-off between performance and resources. Moreover, this information is also represented visually in Figure 2.

GBT-SAM significantly outperforms all other methods in terms of DS on the Adult Glioma dataset (DS_1), achieving 93.54. Furthermore, it accomplishes this with only 6.40 million trainable parameters, the lowest among all evaluated methods. For comparison, MedSegDiff v2 [49] achieves the next highest

Method	# params (M) ↓	DS_1 ↑
nnUNet [12]	30.76	88.50
VTU-Net [39]	17.00	88.73
SegDiff [47]	23.00	85.70
MedSegDiff [48]	25.00	88.90
MedSegDiff v2 [49]	46.00	90.80
SAMed [5]	18.81	77.30
Med-SA [6]	13.00	88.70
GBT-SAM	**6.40**	**93.54**

Table 7: **Efficiency vs Performance.** Comparison of GBT-SAM with state-of-the-art for the DS in Adult Glioma dataset (DS_1) and the number of trainable parameters. This table includes only methods with $< 50M$ trainable parameters. Our method achieves the highest DS while requiring the smallest number of trainable parameters, demonstrating superior efficiency and performance.

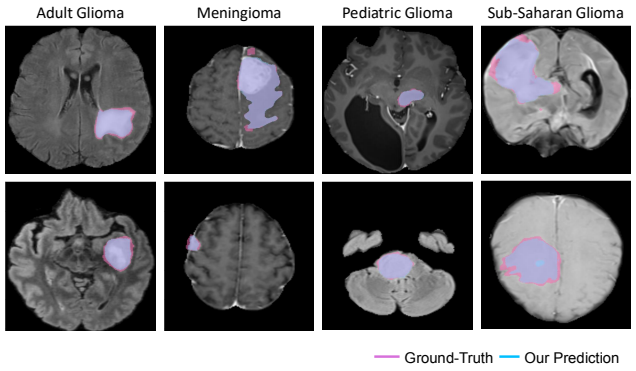

Figure 7: **Visual comparison of ground-truth (pink) and our predictions (blue)**. The purple regions represent the overlap between the ground-truth and our predictions. Each column shows two examples per domain, demonstrating that the predicted segmentations closely match the annotations across diverse cases.

DS (90.80) but requires 46 million parameters, over seven times more than GBT-SAM.

This comparison highlights the superior efficiency of GBT-SAM, achieving the highest segmentation accuracy with the fewest trainable parameters.

5.4. Examples of Qualitative Results

The visual results in Figure 7 show that GBT-SAM produces segmentations that align closely with the ground-truth annotations across all four studied brain tumour domains. The overlap (purple regions) highlights the ability of the model to accurately delineate tumour boundaries and extension. However, there are minor inaccuracies in certain pixels where the model slightly over- or under-segments the tumour regions, particularly in more complex cases or regions with low contrast. These results reflect the robustness of the method in generalizing across different tumour types while acknowledging potential refinement in future work.

5.5. Method Discussion & Limitations

Our method effectively leverages all four MRI modalities (T1, T2, T1c, and T2-FLAIR), enabling strong generalization across brain tumour domains. By adapting SAM with LoRA blocks and using a slice selection strategy, we efficiently train on less than 2.6% of images per iteration (4 selected slices against 155 available ones) without compromising performance. Additionally, the integration of the Depth-Condition block enriches feature representations by capturing inter-slice correlations, a critical factor for volumetric medical data.

Our approach relies on bounding box prompts, designed according to interactive segmentation guidelines. While this ensures strong performance and aligns with clinical use cases, further exploration into unsupervised box prompts could provide a valuable avenue for reducing dependency on annotations.

Despite the demonstrated effectiveness of GBT-SAM, its specialization for brain tumour segmentation introduces certain limitations. The modifications made to the patch embedding layer to process four MRI types simultaneously make the architecture not suitable for volumetric datasets that have a different number of modalities, such as BTCV [77] or Fundus [78], hence being only optimized for 4D brain tumour data.

6. Conclusion

In this work, we propose GBT-SAM, a novel framework for brain tumour segmentation that extends SAM to process multi-parametric MRI (mp-MRI) data. Our approach effectively adapts SAM for the medical imaging domain through key modifications, including a restructured patch embedding layer to accommodate four MRI modalities, parameter-efficient tuning with LoRA blocks, and the integration of a Depth-Condition block to capture inter-slice correlations. These contributions enable our model to achieve state-of-the-art performance on Adult Glioma with a DS of 93.54, while also demonstrating robust generalization to other tumour domains, such as Meningioma, Pediatric Glioma, and Sub-Saharan Glioma.

Our method also achieves significant computational efficiency, utilizing less than 6.5M trainable parameters and processing only 2.6% of slices per iteration without compromising performance. Additionally, the use of bounding box prompts aligns with interactive segmentation guidelines, though future exploration of unsupervised box prompt generation could further enhance applicability.

While optimized for mp-MRI datasets with multiple modalities, GBT-SAM is not directly applicable to datasets with fewer imaging modalities. Nevertheless, our work represents a significant step forward in efficient and generalizable brain tumour segmentation, setting the stage for further advancements.

Acknowledgement. This work has been partially supported by the SEGA-CV (TED2021-131643A-I00) and the HVD (PID2021-125051OB-I00) projects of the Ministerio de Ciencia e Innovación of the Spanish Government.

Privacy Considerations. Since our data originates from an international challenge, all datasets have been anonymized and preprocessed to ensure privacy and eliminate concerns related to image registration [2]. This preprocessing phase maintain the data integrity while adhering to ethical and legal standards.

References

[1] Michael Weller, Wolfgang Wick, Ken Aldape, Michael Brada, Mitchell Berger, Stefan M Pfister, Ryo Nishikawa, Mark Rosenthal, Patrick Y Wen, Roger Stupp, et al. Glioma. *Nature reviews Disease primers*, 1(1):1–18, 2015.

[2] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[3] Jiaqi Wang, Zhengliang Liu, Lin Zhao, Zihao Wu, Chong Ma, Sigang Yu, Haixing Dai, Qiushi Yang, Yiheng Liu, Songyao Zhang, et al. Review of large vision models and visual prompt engineering. *Meta-Radiology*, page 100047, 2023.

[4] Roberto Alcover-Couso, Marcos Escudero-Viñolo, Juan C SanMiguel, and Jesus Bescos. Vlms meet uda: Boosting transferability of open vocabulary segmentation with unsupervised domain adaptation. *arXiv preprint arXiv:2412.09240*, 2024.

[5] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

[6] Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

[7] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[8] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021.

[9] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022.

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

[11] Sheng He, Rina Bao, Jingpeng Li, P Ellen Grant, and Yangming Ou. Accuracy of segment-anything model (sam) in medical image segmentation tasks. *arXiv preprint arXiv:2304.09324*, 2023.

[12] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[13] Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *Medical Image Analysis*, 92:103061, 2024.

[14] Dominic LaBella, Maruf Adewole, Michelle Alonso-Basanta, Talissa Altes, Syed Muhammad Anwar, Ujjwal Baid, Timothy Bergquist, Radhika Bhalerao, Sully Chen, Verena Chung, et al. The asnr-miccai brain tumor segmentation (brats) challenge 2023: Intracranial meningioma. *arXiv preprint arXiv:2305.07642*, 2023.

[15] Anahita Fathi Kazerooni, Nastaran Khalili, Xinyang Liu, Debanjan Haldar, Zhifan Jiang, Syed Muhammed Anwar, Jake Albrecht, Maruf Adewole, Udunna Anazodo, Hannah Anderson, et al. The brain tumor segmentation (brats) challenge 2023: Focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds). *arXiv preprint arXiv:2305.17033*, 2023.

[16] Maruf Adewole, Jeffrey D Rudie, Anu Gbdamosi, Oluyemisi Toyobo, Confidence Raymond, Dong Zhang, Olubukola Omidiji, Rachel Akinola, Mohammad Abba Suwaid, Adaobi Emegoakor, et al. The brain tumor segmentation (brats) challenge 2023: glioma segmentation in sub-saharan africa patient population (brats-africa). *ArXiv*, 2023.

[17] B POPE Whitney and Garth Brandal. Conventional and advanced magnetic resonance imaging in patients with high-grade glioma. *The quarterly journal of nuclear medicine and molecular imaging: official publication of the Italian Association of Nuclear Medicine (AIMN)[and] the International Association of Radiopharmacology (IAR),[and] Section of the Society of...*, 62(3):239, 2018.

[18] Cecilia Diana-Albelda, Roberto Alcover-Couso, Álvaro García-Martín, and Jesus Bescos. How sam perceives different mp-mri brain tumor domains? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4959–4970, 2024.

[19] Tassilo Wald, Saikat Roy, Gregor Koehler, Nico Disch, Maximilian Rouven Rokuss, Julius Holzschuh, David Zimmerer, and Klaus Maier-Hein. Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model. In *Medical Imaging with Deep Learning, short paper track*, 2023.

[20] André Ferreira, Naida Solak, Jianning Li, Philipp Dammann, Jens Kleesiek, Victor Alves, and Jan Egger. How we won brats 2023 adult glioma challenge? just faking it! enhanced synthetic data augmentation and model ensemble for brain tumour segmentation. *arXiv preprint arXiv:2402.17317*, 2024.

[21] James G Nicholson and Howard A Fine. Diffuse glioma heterogeneity and its therapeutic implications. *Cancer discovery*, 11(3):575–590, 2021.

[22] Isabel Esteban-Rodríguez, Samuel López-Muñoz, Luis Blasco-Santana, Jaime Mejías-Bielsa, Carlos H Gordillo, and José A Jiménez-Heffernan. Cytological features of diffuse and circumscribed gliomas. *Cytopathology*, 35(5):534–544, 2024.

[23] Ricky Chen, Matthew Smith-Cohn, Adam L Cohen, and Howard Colman. Glioma subclassifications and their clinical significance. *Neurotherapeutics*, 14(2):284–297, 2017.

[24] HJ Scherer. Cerebral astrocytomas and their derivatives. *The American Journal of Cancer*, 40(2):159–198, 1940.

[25] Pieter Wesseling, Martin van den Bent, and Arie Perry. Oligodendroglioma: pathology, molecular mechanisms and markers. *Acta neuropathologica*, 129:809–827, 2015.

[26] David Maintz, Klaus Fiedler, Jens Koopmann, Britta Rollbrocker, Stefan Nechev, Doris Lenartz, Armin P Stangl, David N Louis, Johannes Schramm, Otmar D Wiestler, et al. Molecular genetic evidence for subtypes of oligoastrocytomas. *Journal of Neuropathology & Experimental Neurology*, 56(10):1098–1104, 1997.

[27] Hans-Georg Wirsching and Michael Weller. Glioblastoma. *Malignant Brain Tumors: State-of-the-Art Treatment*, pages 265–288, 2017.

[28] Aaron C Tan, David M Ashley, Giselle Y López, Michael Malinzak, Henry S Friedman, and Mustafa Khasraw. Management of glioblastoma: State of the art and future directions. *CA: a cancer journal for clinicians*, 70(4):299–312, 2020.

[29] Dietmar Krex, Barbara Klink, Christian Hartmann, Andreas Von Deimling, Torsten Pietsch, Matthias Simon, Michael Sabel, Joachim P Steinbach, Oliver Heese, Guido Reifenberger, et al. Long-term survival with glioblastoma multiforme. *Brain*, 130(10):2596–2606, 2007.

[30] PD Delgado-López and EM Corrales-García. Survival in glioblastoma: a review on the impact of treatment modalities. *Clinical and Translational Oncology*, 18(11):1062–1071, 2016.

[31] Zhifan Jiang, Daniel Capellán-Martín, Abhijeet Parida, Xinyang Liu, María J Ledesma-Carbayo, Syed Muhammad Anwar, and Marius George Linguraru. Enhancing generalizability in brain tumor segmentation: Model ensemble with adaptive post-processing. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2024.

[32] Suraj D Serai. Basics of magnetic resonance imaging and quantitative parameters t1, t2, t2*, t1rho and diffusion-weighted imaging. *Pediatric radiology*, 52(2):217–227, 2022.

[33] K Noguchi, T Ogawa, A Inugami, H Fujita, J Hatazawa, E Shimosegawa, T Okudera, K Uemura, and H Seto. Mri of acute cerebral infarction: a comparison of flair and t2-weighted fast spin-echo imaging. *Neuroradiology*, 39:406–410, 1997.

[34] Hyon Bin Na and Taeghwan Hyeon. Nanostructured t1 mri contrast agents. *Journal of Materials Chemistry*, 19(35):6267–6273, 2009.

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[36] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016.

[37] Rehan Raza, Usama Ijaz Bajwa, Yasar Mehmood, Muhammad Waqas Anwar, and M Hassan Jamal. dresu-net: 3d deep residual u-net based brain tumor segmentation from multimodal mri. *Biomedical Signal Processing and Control*, 79:103861, 2023.

[38] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[39] Himashi Peiris, Munawar Hayat, Zhaolin Chen, Gary Egan, and Mehrtash Harandi. A robust volumetric transformer for accurate 3d tumor segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 162–172. Springer, 2022.

[40] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[41] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[42] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[43] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021.

[44] Marjan Vatanpour and Javad Haddadnia. Transdoubleu-net: Dual scale swin transformer with dual level decoder for 3d multimodal brain tumor segmentation. *IEEE Access*, 11:125511–125518, 2023.

[45] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[46] Pablo Marcos-Manchón, Roberto Alcover-Couso, Juan C SanMiguel, and Jose M Martínez. Open-vocabulary attention maps with token optimization for semantic segmentation in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9242–9252, 2024.

[47] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021.

[48] Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning*, pages 1623–1639. PMLR, 2024.

[49] Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6030–6038, 2024.

[50] Anqi Xiao, Keyi Han, Xiaojing Shi, Jie Tian, and Zhenhua Hu. Data augmentation with multi-armed bandit on image deformations improves fluorescence glioma boundary recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 130–140. Springer, 2024.

[51] Roberto Alcover-Couso, Juan C SanMiguel, Marcos Escudero-Vinolo, and Pablo Carballeira. Per-class curriculum for unsupervised domain adaptation in semantic segmentation. *The Visual Computer*, pages 1–19, 2024.

[52] Roberto Alcover-Couso, Marcos Escudero-Viñolo, Juan C SanMiguel, and Jesus Bescós. Gradient-based class weighting for unsupervised domain adaptation in dense prediction visual tasks. *arXiv preprint arXiv:2407.01327*, 2024.

[53] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023.

[54] Guoyao Deng, Ke Zou, Kai Ren, Meng Wang, Xuedong Yuan, Sancong Ying, and Huazhu Fu. Sam-u: Multi-box prompts triggered uncertainty estimation for reliable sam in medical image. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 368–377. Springer, 2023.

[55] Jiayuan Zhu, Yunli Qi, and Junde Wu. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*, 2024.

[56] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[57] Suresh Samudrala and C Krishna Mohan. Semantic segmentation of breast cancer images using densenet with proposed pspnet. *Multimedia Tools and Applications*, 83(15):46037–46063, 2024.

[58] Lareib Fatima Talib, Javaria Amin, Muhammad Sharif, and Mudassar Raza. Transformer-based semantic segmentation and cnn network for detection of histopathological lung cancer. *Biomedical Signal Processing and Control*, 92:106106, 2024.

[59] Reza Azad, Moein Heidary, Kadir Yilmaz, Michael Hüttemann, Sanaz Karimijafarbigloo, Yuli Wu, Anke Schmeink, and Dorit Merhof. Loss functions in the era of semantic segmentation: A survey and outlook. *arXiv preprint arXiv:2312.05391*, 2023.

[60] 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. Segment anything, 2023.

[61] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[62] Yang Liu, Yao Zhang, Yixin Wang, Feng Hou, Jin Yuan, Jiang Tian, Yang Zhang, Zhongchao Shi, Jianping Fan, and Zhiqiang He. A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[63] Yimu Pan, Sitao Zhang, Alison D Gernand, Jeffery A Goldstein, and James Z Wang. Ai-sam: Automatic and interactive segment anything model. *arXiv preprint arXiv:2312.03119*, 2023.

[64] Tongxue Zhou, Su Ruan, and Haigen Hu. A literature survey of mr-based brain tumor segmentation with missing modalities. *Computerized Medical Imaging and Graphics*, 104:102167, 2023.

[65] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

[66] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[67] Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. *arXiv preprint arXiv:2310.17513*, 2023.

[68] Yucheng Tang, Jie Liu, Zongwei Zhou, Xin Yu, and Yuankai Huo. Efficient 3d representation learning for medical image analysis. In *Deep Learning For 3d Vision: Algorithms And Applications*, pages 399–437. 2024.

[69] Yu-Cheng Chou, Zongwei Zhou, and Alan Yuille. Embracing massive medical data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 24–35. Springer, 2024.

[70] Gaoyu Xiao, B Nicolas Bloch, Jonathan Chappelow, Elizabeth M Genega, Neil M Rofsky, Robert E Lenkinski, John Tomaszewski, Michael D Feldman, Mark Rosen, and Anant Madabhushi. Determining histology-mri slice correspondences for defining mri-based disease signatures of prostate cancer. *Computerized Medical Imaging and Graphics*, 35(7-8):568–578, 2011.

[71] Kihwan Choi, Joon Seok Lim, and Sungwon Kim. Self-supervised inter- and intra-slice correlation learning for low-dose ct image restoration without ground truth. *Expert Systems with Applications*, 209:118072, 2022.

[72] Minhyeok Lee. Gelu activation function in deep learning: a comprehensive mathematical analysis and performance. *arXiv preprint arXiv:2305.12073*, 2023.

[73] Jeroen Bertels, Tom Eelbode, Maxim Berman, Dirk Vandermeulen, Frederik Maes, Raf Bisschops, and Matthew B Blaschko. Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 92–100. Springer, 2019.

[74] Tom Eelbode, Jeroen Bertels, Maxim Berman, Dirk Vandermeulen, Frederik Maes, Raf Bisschops, and Matthew B Blaschko. Optimization for

medical image segmentation: theory and practice when evaluating with dice score or jaccard index. *IEEE transactions on medical imaging*, 39(11):3679–3690, 2020.

[75] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[76] Jay N Paranjape, Shameema Sikder, S Swaroop Vedula, and Vishal M Patel. Low-rank adaptation of segment anything model for surgical scene segmentation. In *International Conference on Pattern Recognition*, pages 187–202. Springer, 2025.

[77] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, page 12, 2015.

[78] N Balakrishna, MB Mukesh Krishnan, E Venkata Ram Sai, S Vinish Ranganath, K Sonika, and L Gouri Priyanka. Ocular disease recognition using efficientnet. In *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 1069–1074. IEEE, 2024.