Brain Metastasis Segmentation on Pretreatment MRI:

A Comparative Study of Volumetric SwinUNETR and Slice-Based YOLOv8 Pipelines

A thesis submitted in partial fulfillment of the requirements for Honors in the Department of Computer Science at Brown University

Shruti Panse

Ritambhara Singh, Ph.D *(Advisor)*
Zhicheng Jiao, Ph.D *(Second Reader)*

BROWN

April 2026

Acknowledgements

My deepest gratitude goes to Professor Ritambhara Singh for advising me throughout this project. She took me into her lab and gave me the space to figure things out on my own, while being there whenever I needed guidance, and I learned a lot from her about how to approach research. I'm also grateful to Professor Zhicheng Jiao for serving as my second reader. Jack Roberts, a masters student in the Singh Lab, has been a constant resource throughout the year, helping me think through research ideas and keeping the scope of the project manageable when I wasn't sure where to take it next. I am also grateful to Dr. Reshma Munbodh and Dr. Ziqing Shi, who I met with monthly throughout the year and whose ideas and clinical context helped shape the direction of this work. I thank the Brown University Center for Computation and Visualization for access to the Oscar cluster, and the Yale School of Medicine, The Cancer Imaging Archive, and the BraTS organizers for making the datasets used in this thesis publicly available. Finally, I am incredibly grateful to my mom, my dad, my brother Rohit, and my friends for their endless support and encouragement throughout my time at Brown. I would not have been able to complete this thesis without them.

Brain Metastasis Segmentation on Pretreatment MRI: A Comparative Study of Volumetric SwinUNETR and Slice-Based YOLOv8 Pipelines

Shruti Panse
Ritambhara Singh, Ph.D
Zhicheng Jiao, Ph.D

Abstract

Brain metastases are the most common malignant intracranial tumors in adults, occurring in roughly 10–30% of cancer patients over the course of their disease. Accurate segmentation of these lesions on MRI is essential for treatment planning and longitudinal response assessment, but manual delineation is time-consuming and produces substantial inter-rater variability, which has motivated a growing body of work applying deep learning to this task. Even with this growing popularity, however, gaps remain in this body of work, two of which this thesis focuses on. (1) Most studies report results on a single dataset and do not isolate how individual methodological choices, such as input dimensionality, modality coverage, training duration contribute to performance. (2) Few studies provide a side-by-side comparison of volumetric and slice-based architectures evaluated on the same brain metastasis data under a consistent protocol. To address these gaps, this thesis presents a comparative study organized into three phases: pipeline verification through smoke and BraTS sanity tests, development of a slice-based YOLOv8 comparator, and adaptation of a 3D SwinUNETR pipeline to the Pretreat-MetsToBrain-Masks dataset. The BraTS sanity run reproduced published-level performance (mean Dice 0.908), confirming the volumetric implementation. The YOLOv8 comparator reached an overall lesion Dice of 0.896 after a multiclass formulation fix and the addition of 2.5D context, showing that slice-based methods can be competitive when carefully tuned. SwinUNETR on Pretreat reached a best validation Dice of 0.774, with substantial gains coming separately from richer input channels (single-channel T1c to four-channel multimodal, 0.607 to 0.761) and from longer training with self-supervised initialization (0.774 in the single-channel SSL configuration), together producing the main volumetric per-class segmentation result of the thesis. The two model families produce different kinds of output and are evaluated under different metrics, and are therefore best read as complementary rather than directly comparable: SwinUNETR produces volumetric per-patient segmentations, while YOLO produces pixel-pooled slice-level predictions. Across both pipelines, the largest gains came from input representation, task formulation, and training duration rather than from incremental architectural refinements. Extending these methods to large unlabeled cohorts, including the Yale-Brain-Mets-Longitudinal collection, is left as future work.

Contents

1 Introduction

Brain metastases are the most common malignant intracranial tumors in adults[1]. They occur in roughly 10–30% of cancer patients over the course of their disease, with the majority of cases originating from lung, breast, or melanoma primaries[1,2]. They typically present as small, multifocal enhancing lesions scattered throughout the brain, and identifying exactly where each lesion sits matters a great deal for how a patient is treated[2,3]. The standard treatment for many patients is stereotactic radiosurgery, which requires every individual lesion to be carefully outlined on contrast-enhanced MRI before the radiation plan can be calculated[4]. In current clinical practice, this delineation is performed by hand by a neuroradiologist or radiation oncologist, slice by slice, lesion by lesion[5].

This is a slow process, and it is far from perfectly reproducible. Different clinicians draw the same lesion differently, and the smaller the lesion, the more disagreement there tends to be — which is particularly unfortunate, since the smallest lesions are also the ones most likely to benefit from early radiosurgical treatment[4–7]. Deep learning has accordingly emerged as a natural fit for automating brain tumor segmentation, and the literature in this area has matured substantially over the past decade[8,9]. Much of the architectural progress has happened in the context of the BraTS challenge series[6,10], which has produced everything from U-Net derivatives to transformer-based volumetric models such as SwinUNETR[11]. Most of this progress, however, has happened on the glioma task, where lesions tend to be larger and more contiguous than the small, disseminated metastases that show up in real clinical practice[12]. Methods that perform well on glioma benchmarks do not always transfer cleanly to the brain metastasis setting, and the field has only recently begun to converge on dedicated metastasis benchmarks such as BraTS-METS[6,10].

In this study, I focus on two specific gaps in the existing literature. (1) Most published studies report a single model evaluated on a single dataset, which makes it difficult to tell which methodological choices, such as input dimensionality, modality coverage, or training duration, are actually driving the performance[12]. (2) Few studies compare qualitatively different architectures, such as volumetric and slice-based models, on the same data under a consistent protocol. To address these gaps, I present a comparative study organized into three phases: pipeline verification through smoke and BraTS sanity tests, development of a slice-based YOLOv8 comparator, and in-domain adaptation of a 3D SwinUNETR pipeline on the Pretreat-MetsToBrainMasks dataset[13] as the main volumetric segmentation result.

2 Background

2.1 Brain Metastases: Clinical Context

Brain metastases form when cancer cells from a primary tumor elsewhere in the body travel through the bloodstream and seed in the brain parenchyma[2]. They are the most common malignant intracranial tumors in adults, far outnumbering primary brain tumors, with non-small cell lung cancer, breast cancer, and melanoma accounting for the majority of cases[1,2]. The clinical course is shaped by the number, size, and location of lesions, and the standard of care has shifted over the past decade from whole-brain radiotherapy toward more targeted modalities such as stereotactic radiosurgery (SRS), which requires accurate per-lesion delineation[4,14]. Treatment planning therefore depends on a high-quality segmentation of each individual metastasis on contrast-enhanced MRI, and the time spent on manual delineation is an explicit bottleneck in the radiation oncology workflow[7,15].

A multi-center randomized study by Luo and colleagues evaluated AI-assisted versus manual delineation in a stereotactic radiosurgery workflow, reporting roughly 42% time savings when AI-generated contours were used as a starting point for clinician review without measurable

loss of segmentation quality[7]. The clinical motivation for automated brain metastasis segmentation is therefore not in doubt; the open question, which this thesis addresses, is which modeling choices most affect segmentation quality on a representative pretreatment dataset.

2.2 Magnetic Resonance Imaging for Brain Tumors

Contrast-enhanced MRI is the standard imaging modality for brain metastasis detection and treatment planning. The four sequences acquired in nearly every brain tumor protocol are T1-weighted (T1), T1-weighted post-gadolinium contrast (T1c), T2-weighted (T2), and fluid-attenuated inversion recovery (FLAIR), each highlighting different aspects of the tumor and surrounding brain[3]. Contrast-enhanced T1 is the workhorse of metastasis detection: enhancing tumor tissue appears bright relative to surrounding parenchyma because of breakdown of the blood-brain barrier in the tumor region[3]. FLAIR is the natural sequence for visualizing peritumoral edema and the broader extent of tumor-associated tissue change, while T2 provides additional contrast for non-enhancing tumor components and edema[13,16]. Figure 1 shows the four sequences for a representative patient.

Segmentation tasks defined on this multimodal substrate have generally targeted three distinct sub-regions: the contrast-enhancing tumor, the necrotic or non-enhancing tumor core, and the surrounding peritumoral edema[6]. The clinical importance of each sub-region differs by treatment modality: stereotactic radiosurgery treatment plans are usually drawn around the enhancing component[4], while surgical planning and response assessment may consider the full tumor volume[16]. The Pretreat-MetsToBrain-Masks dataset used in this thesis provides 3D segmentations for all three sub-regions across 200 patients with brain metastases[13].

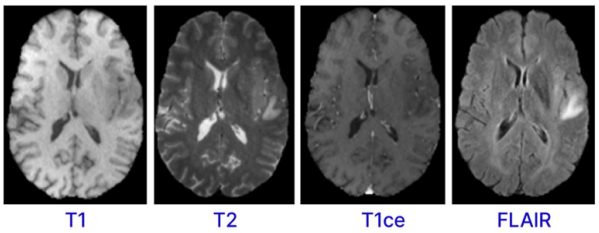

Figure 1: The four standard MRI sequences acquired in brain tumor imaging protocols, shown for the same patient. Each sequence highlights different aspects of the tumor: T1 provides anatomical reference, T1c (post-gadolinium contrast) highlights the contrast-enhancing tumor where the blood-brain barrier has broken down, T2 provides additional contrast for non-enhancing tumor components, and FLAIR is the natural sequence for visualizing peritumoral edema. No single sequence captures all aspects of the tumor; the four sequences together provide complementary information used by the multimodal segmentation pipelines in this thesis.

2.3 Deep Learning for Medical Image Segmentation

Modern medical image segmentation is dominated by deep learning, and the architectural lineage of the field traces back to the U-Net of Ronneberger and colleagues[17]. The U-Net's encoder-decoder structure with skip connections has become the default starting point for biomedical segmentation, and almost every subsequent architecture either extends or competes with it[8].

Subsequent work has pursued two main axes of improvement. The first is volumetric processing: medical images are typically three-dimensional, and 3D convolutional architectures such as 3D U-Net and DeepMedic exploit through-plane context that 2D networks cannot access[8]. The second is the introduction of attention mechanisms and transformer-based architectures, which have moved from natural language processing into computer vision and now into medical imaging[18]. SwinUNETR, which combines a Swin transformer encoder with a U-Net-style decoder, is one prominent example of this trend and is the volumetric architecture used in this thesis[11].

A parallel line of work, motivated less by

architectural novelty than by deployment efficiency, has applied detection-and-segmentation pipelines such as the YOLO family to medical slices[8]. These models operate on 2D slices and predict bounding boxes plus per-pixel masks for each detected object, and although they sacrifice volumetric context they are fast, lightweight, and benefit from the large body of pretrained weights available for natural-image tasks. The slice-based comparator in this thesis is built on YOLOv8, configured for multiclass segmentation of contrast-enhancing tumor, necrotic core, and peritumoral edema.

2.4 Evaluation Metrics

Segmentation quality in this thesis is reported primarily as the Dice similarity coefficient, defined as twice the size of the intersection between predicted and ground-truth masks divided by the sum of their sizes. Dice ranges from 0 (no overlap) to 1 (perfect overlap) and is the metric used in the BraTS challenge series and most of the brain tumor segmentation literature[6,10]. Dice can be computed at different granularities: per patient (averaged across the validation cohort), per region (separately for tumor sub-regions), or globally over all evaluated pixels and slices[8]. The choice matters for cross-method comparison, as discussed in Section 4.6.

Detection-oriented metrics are also relevant for the YOLO pipeline. Mean average precision at IoU thresholds of 0.5 and 0.5–0.95 (mAP@0.5 and mAP@0.5:0.95) capture the quality of bounding-box and mask proposals, and precision and recall provide complementary views of the detector's behavior. Standard reviews emphasize that lesion-wise sensitivity and Dice should be reported together for brain metastasis tasks because lesion size varies dramatically within and across patients[12].

3 Related Works

Brain tumor segmentation in MRI is one of the most active areas of medical image analysis, and the literature is too large to review exhaustively. The summary that follows is therefore organized around the methodological choices that bear directly on the experiments in Section 5: the U-Net family of convolutional architectures, the transformer-based volumetric models from which SwinUNETR is drawn, the brain metastasis-specific approaches that motivate the slice-based YOLOv8 detection-and-segmentation pipeline used here, and the public datasets on which all of this work is built.

3.1 Convolutional Architectures and the U-Net Family

The U-Net is the foundational architecture for medical image segmentation, and almost every modern segmentation model can trace some part of its design back to it[17]. Its encoder-decoder structure with skip connections has held up well across modalities and anatomical regions, and three-dimensional extensions such as 3D U-Net and V-Net dominated the BraTS challenge for several years[8]. Subsequent work has refined rather than replaced this template. The nnU-Net framework, introduced by Isensee and colleagues, automates pre-processing, hyperparameter selection, and ensembling around a standard U-Net backbone, and has proven difficult to beat on a wide range of segmentation benchmarks, including brain tumor tasks[19]. These architectures have been evaluated extensively on the BraTS glioma data, where the best models now exceed Dice scores of 0.9 for the whole tumor[11,19]. Performance on out-of-distribution clinical data has been studied less. Berkley and colleagues, working in the Singh and Munbodh laboratories at Brown, addressed this question directly by training a state-of-the-art 3D U-Net on BraTS and evaluating it on in-house clinical MRIs that differed from BraTS in resolution, standardization, and tumor type, reporting Dice scores of 0.764, 0.648, and 0.610 for whole tumor, tumor core, and enhancing tumor respectively, with no statistically significant difference from inter-annotation variability between two expert clinical radiation oncologists[5]. Their conclusion is that modern segmentation models transfer well to new clinical institutions, considerably improving on the pre-

vious generation. I take two things from this literature into the present work. First, the U-Net family sets the implicit baseline that any new approach has to clear, and the BraTS sanity-check experiment in Section 5 uses a tumor-region setup of this kind to confirm that the training pipeline is functioning correctly before the brain metastasis-specific results are interpreted. Second, the lower clinical numbers compared with the BraTS validation set are a useful reminder that benchmark performance and clinical performance are measured on different distributions, which is part of why this thesis evaluates more than one architecture on the brain metastasis dataset rather than relying on a single model and a single benchmark.

3.2 Transformer-Based Volumetric Segmentation

Transformers entered medical image segmentation as a response to a known weakness of pure convolutional networks: their limited receptive field. UNETR was the first model to apply a vision transformer encoder to volumetric medical data, treating a 3D volume as a sequence of patches and connecting the transformer to a convolutional decoder via skip connections[18]. SwinUNETR refined this idea by replacing the plain transformer with a Swin transformer, whose shifted-window attention scheme makes self-attention practical at multiple resolutions, and pairing it with a U-Net-style decoder so that the model captures both fine local detail and long-range structural context[11]. On the BraTS 2021 challenge, SwinUNETR was among the top-performing approaches in the validation phase[11]. Subsequent hybrids such as TransBTS combine 3D convolutional encoders with transformer bottlenecks for global feature modeling, with each new variant claiming small improvements on the BraTS benchmark[20].

What I take from this literature is not just a model choice but a training recipe: SwinUNETR benefits from pretraining on large unlabeled medical-image corpora, and the experiments in Section 5 take advantage of this property by initializing the model from publicly released self-supervised weights[21] before fine-tuning on the brain metastasis dataset. This setup makes SwinUNETR the primary volumetric architecture studied in this thesis, against which a slice-based detection-and-segmentation pipeline is then compared.

3.3 Brain Metastasis-Specific Approaches

General-purpose tumor segmentation architectures often perform well on glioma data but struggle on brain metastases, which are smaller, more numerous, and more sparsely distributed than the diffuse gliomas that the BraTS benchmark was originally built around. A meta-analysis of 37 studies found that deep learning models achieve a pooled lesion-wise Dice score of approximately 0.79 on brain metastasis tasks, with patient- and lesion-wise sensitivities near 0.86 and 0.87 respectively, and that U-Net-based methods consistently outperform alternatives[12]. The same review identified MRI hardware diversity, slice thickness, and the choice of input sequences as significant factors affecting detection sensitivity[12].

Two themes recur across the metastasis-specific literature. The first is the use of two-stage detection-then-segmentation pipelines. Because metastases are small and sparse relative to the brain volume, training a single segmentation model on full volumes wastes capacity on background voxels. Pipelines such as DeSeg, introduced by Yu and colleagues, instead use a detection module to localize candidate lesions and then perform detailed segmentation on cropped regions of interest, achieving a sensitivity of 0.91 and a positive predictive value of 0.77 on small lesions, and a Dice score of 0.86 on large ones[22]. The second is the use of longitudinal information: Huang and colleagues introduced DeepMedic+, in which the segmentation mask from a prior visit is fed to the network as an additional input pathway alongside a volume-level sensitivity-specificity loss[23], and Patel and colleagues extended this idea further with SPIRS, a joint registration-and-segmentation network that warps prior annota-

tions onto a new time point, with reported improvements particularly for micro-metastases[24].

The slice-based YOLOv8 pipeline used in Section 5 follows the detection-then-segmentation philosophy of DeSeg, applied at the level of two-dimensional axial slices: a detection head proposes candidate lesion regions on each slice, and a segmentation head then refines those proposals into per-pixel masks. Working slice-by-slice rather than over full volumes makes the pipeline computationally cheap and lets it borrow from the large body of pretrained weights available for two-dimensional natural-image tasks. The longitudinal methods are not directly used in this thesis, since the experiments are conducted on a single-time-point pretreatment dataset, but the longitudinal setting is central to the future-work discussion in Section 6.5.

Beyond detection-segmentation cascades and longitudinal priors, brain metastasis segmentation has also benefited from approaches that exploit spatial structure beyond the regular voxel grid. Saueressig and colleagues, working in the Munbodh and Singh laboratories at Brown, proposed a joint graph and image convolution network for the BraTS 2021 challenge that models each brain as a graph composed of distinct image regions, where a graph neural network produces an initial segmentation that captures global feature interactions and a voxel-level convolutional network then refines the result using local image detail[25]. What I take from this work is its underlying question — how to combine global structural information with local intensity detail — which sits at the heart of the choice between the volumetric SwinUNETR architecture and the slice-based YOLOv8 pipeline studied in Section 5.

3.4 Public Datasets and Benchmarks

The progress documented in the previous subsections has depended critically on the availability of public datasets. The BraTS challenge series has been the central benchmark for over a decade, expanding from glioma segmentation to include pediatric tumors, meningiomas, and brain metastases in recent years[8,10]. The BraTS-METS 2023 challenge, organized by Moawad and colleagues, established the first standardized brain metastasis segmentation benchmark by aggregating retrospectively collected pre-treatment multiparametric MRI scans from eight international institutions and annotating them in a four-stage workflow involving algorithm-assisted student annotators, neuroradiologist review, and final approval by a board-certified neuroradiologist[6]. The 2025 Lighthouse challenge extended this work by formalizing the inter- and intra-rater variability of the annotation process and by including post-treatment cases[10].

The Pretreat-MetsToBrain-Masks dataset, a 200-patient resource of annotated pretreatment brain metastasis MRI, was developed at Yale to provide a heterogeneous and well-curated collection for training and validation of clinical segmentation algorithms[13]. Its primary-tumor distribution roughly mirrors the clinical epidemiology of brain metastasis, with non-small cell lung cancer most common and melanoma, breast cancer, small cell lung cancer, renal cell carcinoma, and gastrointestinal cancers each represented in smaller numbers[13]. All four standard MRI sequences are provided, with segmentations on T1 post-contrast covering the contrast-enhancing tumor and necrosis and FLAIR-based segmentations covering peritumoral edema[13]. This is the dataset on which both the SwinUNETR and YOLOv8 experiments in Section 5 are conducted; the specific data partitioning used by each pipeline, and the comparability caveats it introduces, are described in Section 4.

The Yale-Brain-Mets-Longitudinal collection extends this resource into the longitudinal domain. It comprises 11,884 MRI studies from 1,430 patients with clinically confirmed brain metastases, spanning nearly two decades of acquisition, with all four standard sequences and clinical metadata included[15]. According to the dataset authors, it is the largest publicly available longitudinal brain metastasis MRI collection. Unlike Pretreat-MetsToBrain-Masks, however, it does not ship with segmentation la-

bels, which makes it an attractive target for inference with models trained on labeled data but not directly suitable for supervised training without additional annotation. It is in this future-work setting, discussed in Section 6.5, that the Yale-Brain-Mets-Longitudinal collection becomes most relevant to the present thesis.

4 Methods

This section describes the datasets used in the thesis, the preprocessing pipeline that converts raw NIfTI MRI volumes into tensors suitable for the model families, and the three-phase experimental design that organizes all runs reported in Section 5. The phases proceed from pipeline verification (Phase A), through development of a strong slice-based comparator with YOLO (Phase B), to in-domain adaptation of the volumetric SwinUNETR pipeline as the main result (Phase C). Figure 2 summarizes the three-phase design and the headline result from each phase. A larger longitudinal cohort, Yale-Brain-Mets-Longitudinal, was assembled and processed during the project but ultimately reserved for future work; it is described briefly here for completeness and is discussed further in Section 6.5. All code used to produce the experiments below is publicly available at `https://github.com/ShrutiVPanse/BrainSeg`.

4.1 Datasets

Three datasets play distinct roles in this thesis: the BraTS 2021 multimodal training set serves as a controlled sanity benchmark to verify that the volumetric pipeline behaves correctly, the Pretreat-MetsToBrain-Masks dataset is the core in-domain training and validation resource on which the main empirical claims rest, and the Yale-Brain-Mets-Longitudinal collection provides external deployment-style inference evidence on a much larger and more heterogeneous cohort. The slice-based YOLO comparator uses several distinct data formulations derived from Pretreat-MetsToBrain-Masks; these are described in Section 4 as part of the preprocessing pipeline.

BraTS 2021 (Sanity Benchmark)

BraTS 2021 serves as a controlled sanity benchmark to verify that the volumetric pipeline behaves correctly before any in-domain modeling claim is made. The sanity check used the BraTS 2021 multimodal training set under the standard MONAI five-fold cross-validation split (`brats21_folds.json`, fold 1 holdout, 250 cases). BraTS provides a uniformly preprocessed multimodal cohort of glioma cases with the same four standard sequences and with multi-region annotations for tumor core, whole tumor, and enhancing tumor[8]. Although BraTS gliomas differ pathologically from brain metastases, they share the same imaging substrate and label conventions, making them suitable for verifying that the model and training pipeline could reproduce competitive published results. The fold-1 holdout was paired with a locally stored `fold1_model.pt` checkpoint trained on the remaining four folds.

Pretreat-MetsToBrain-Masks (Core In-Domain Dataset)

This is the core in-domain dataset for the main empirical claims of the thesis. The Pretreat-MetsToBrain-Masks dataset is a heterogeneous, annotated collection of pretreatment brain MRI from 200 patients with brain metastasis treated at Yale New Haven Hospital between 2013 and 2021[13]. The cohort is dominated by non-small cell lung cancer, with melanoma, breast cancer, small cell lung cancer, renal cell carcinoma, and gastrointestinal cancers also represented[13]. Each patient has all four standard MRI sequences (T1 pre-contrast, T1 post-contrast, T2, and FLAIR) without significant motion artifact, together with 3D segmentation masks for the contrast-enhancing tumor (CET), necrotic nonenhancing tumor core (NT), and peritumoral edema (PE)[13]. The original dataset provides contrast-enhancing and necrotic segmentations

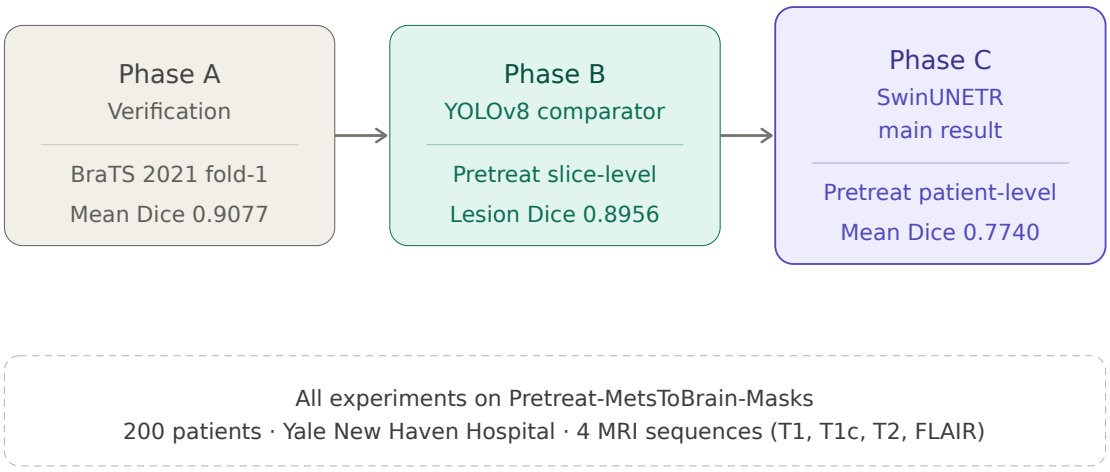

Figure 2: Three-phase experimental design. Phase A verifies the volumetric pipeline reproduces published BraTS performance (mean Dice 0.9077). Phase B develops a slice-based YOLOv8 comparator on Pretreat-MetsToBrain-Masks (lesion Dice 0.8956 on the slice-pooled metric). Phase C adapts SwinUNETR to the Pretreat-MetsToBrain-Masks dataset and constitutes the main empirical contribution of the thesis (mean foreground Dice 0.7740 on a 20-case patient-level holdout). The cross-family comparability caveats are discussed in Section 4.6.

on T1 post-contrast and a whole-tumor segmentation (including peritumoral edema) on FLAIR[13]; in this thesis the labels are processed into a three-class scheme with class values 1 for CET, 2 for NT, and 3 for PE, where PE is treated as edema outside the CET and NT regions. The dataset was used in two manifest formulations: a single-channel formulation supplying only the contrast-enhanced T1 image (`pretreat_t1c_seg_manifest.jsonl`), and a four-channel formulation supplying all four sequences as input channels (`pretreat_4ch_seg_manifest.jsonl`).

Yale-Brain-Mets-Longitudinal (Deployment-Style Inference Evidence)

The Yale-Brain-Mets-Longitudinal collection provides external deployment-style inference evidence on a much larger and more heterogeneous cohort than Pretreat. It is an 11,884-study longitudinal MRI archive from 1,430 patients with confirmed brain metastases[15]. It was assembled and indexed during this project, and an 898-case post-treatment subset (`yale_post_subset_manifest.jsonl`) was prepared for downstream inference. The collection ships pre-skull-stripped from TCIA and was consumed in this thesis as provided, without re-running brain extraction. Because the collection does not include segmentation labels, it was used in this thesis only for qualitative deployment-style inference and coverage analysis (empty-mask rate, qualitative realism) rather than for supervised training or quantitative evaluation. The full deployment-style inference work that this dataset enables is discussed as future work in Section 6.5.

4.2 Preprocessing Pipeline and Compute Environment

The preprocessing pipeline was implemented in Python using NumPy, NiBabel, and MONAI, with PIL/Pillow used for the slice-export step that produces PNG inputs to YOLO. For the SwinUNETR pipeline, no slice-level conversion was performed; the model operated directly

on the original three-dimensional NIfTI volumes loaded through MONAI's data utilities, with patch-based sampling handled by the dataloader. For the YOLO pipeline, a series of scripts converted the volumes into 2.5D image-mask pairs. The `make_slices_2_5d_12ch.py` script extracted, for each in-plane location, a stack of three adjacent axial slices ($z-1$, z, $z+1$) across all four MRI modalities (T1, T2, T1c, FLAIR), producing twelve-channel inputs per sample. A subsequent script (`prepare_yolo_2_5d_dataset.py`) selected three of these twelve channels, defaulting to indices 2, 6, and 10 corresponding to T1c at $z-1$, z, and $z+1$, to populate the RGB inputs expected by YOLO's pretrained backbones. Per-slice polygon labels were then generated from the 3D ground-truth masks via standard contour extraction, with multiclass label values preserved (rather than binarized) to support the three-class formulation.

Three YOLO data formulations were used. The earliest, single-class formulation (`brainmets.yaml`) treated each axial slice as an independent input image with a single foreground class ("metastasis") and was used in the `train` and `train2` baseline runs. The 2D multiclass formulation (`brainmets_yolo_2d_multiclass/brainmets_2d_multiclass.yaml`) extended the input scheme with three foreground classes corresponding to CET, NT, and PE. The 2.5D multiclass formulation (`brainmets_yolo_2_5d/brainmets_2_5d.yaml`) used the three-slice T1c stack described above, retaining the three-class CET/NT/PE label scheme but adding limited through-plane context. Image registration across the four modalities was assumed to have been performed by the dataset providers; BraTS ships with co-registered sequences by construction, and the Pretreat-MetsToBrain-Masks release on TCIA provides the four sequences in a form suitable for joint segmentation.

All training and evaluation were performed on the Brown University Oscar high-performance computing cluster. Different runs were scheduled on different GPU types as available, including NVIDIA Quadro RTX 6000, RTX A5000, RTX A5500, TITAN RTX, GeForce RTX 3090, and L40S accelerators. The software stack used PyTorch 2.9.1 with CUDA 12.9, Python 3.10.19, MONAI for medical-imaging-specific data utilities, and Ultralytics 8.3.230 for the YOLO experiments. SwinUNETR was trained for 100 or 150 epochs depending on the run, using the AdamW optimizer with a base learning rate of 10^{-4}, weight decay of 10^{-5}, batch size of 1, and a fixed $96 \times 96 \times 96$ voxel patch size; no learning-rate schedule and no mixed-precision training were used. The training objective was MONAI's DiceCELoss with one-hot encoding and softmax activation, combining Dice and cross-entropy losses in a single term. YOLO runs used the Ultralytics framework with `optimizer=auto`, which selected AdamW (with learning rates near 0.001 to 0.002) for the 50-epoch GPU runs and SGD (with a learning rate of 0.01) for the 100-epoch runs; image size was 512 throughout. Random seeds were fixed at 42 for the SwinUNETR experiments and 0 for the YOLO experiments to support reproducibility.

4.3 Phase A: Pipeline Verification (Smoke and Sanity Tests)

Phase A consisted of small-scale runs whose purpose was to verify, before any modeling claim was made, that the data, training, and inference pipelines for both model families functioned end to end. Three categories of verification are reported.

The first, `sanity_smoke_ssl`, was a smoke sanity run with self-supervised pretrained SwinViT weights evaluated on a deliberately tiny BraTS slice of two cases. Its purpose was to confirm that the BraTS data could be loaded, that the pretrained weights could be applied, and that the evaluation script ran to completion; a low Dice score was expected and accepted in this configuration. The second, `swinunetr_brats21_fold1_sanity`, was a full fold-1 sanity evaluation using the local `fold1_model.pt` checkpoint on the 250-case validation set. This was a stronger

sanity check designed to confirm that the SwinUNETR implementation, when applied with appropriate trained weights to a curated benchmark, could reproduce published-level performance. The third pair of verification runs, `yolo_2_5d_smoketest` and `yolo_2_5d_smoketest2`, were one-epoch runs on a tiny smoketest dataset to verify that the 2.5D formulation, the YOLO label conversion, and the Dice export script functioned without crashing; meaningful metrics were not expected from a single epoch on four images.

Phase A established the precondition for the rest of the thesis: that subsequent in-domain results would reflect modeling decisions rather than implementation bugs.

4.4 Phase B: YOLO-Based 2D and 2.5D Detection-and-Segmentation

Phase B developed a slice-based comparator using the YOLOv8 family of detection-and-segmentation models, specifically the `yolov8s-seg` variant, accessed through the Ultralytics framework. The phase proceeded through several iterations as instabilities and class-formulation bugs were diagnosed and corrected, ending with a multiclass 2.5D configuration that achieved the highest overall lesion Dice in this thesis.

The earliest baselines were the `train` and `train2` configurations, which used the `brainmets.yaml` dataset definition with a single foreground class ("metastasis") and `yolov8s-seg` pretrained weights. The `train` configuration used CPU and a batch size of 4 for 50 epochs at `imgsz` 512; `train2` used a single GPU with a batch size of 8 for the same number of epochs. These early runs verified end-to-end functionality but exposed instability in cross-class behavior in the exported Dice metrics, with one early GPU run yielding an overall lesion Dice of only 0.21 and per-class scores that were obviously inconsistent with the qualitative outputs.

The pipeline was then shifted to the 2.5D formulation described in Section 4.2, on the hypothesis that limited through-plane context was needed to stabilize multi-class predictions. The `yolo_2_5d_full` configuration ran this setup on CPU for 50 epochs and reached a substantially higher overall lesion Dice of 0.7304. The `yolo_2_5d_gpu` configuration extended this to GPU and reached 0.8761, but inspection of its exported per-class Dice table revealed a class-formulation mismatch: the underlying YOLO model had been trained with a binary class head (`nc=1`) while evaluation expected the three-class Pretreat label scheme, so the CET signal was retained while NT and PE collapsed to near zero. A multiclass fix was introduced in `yolo_2_5d_gpu_multiclass_fix`, which corrected the class formulation in both training and evaluation by enforcing the three-class Pretreat scheme and preserving multiclass labels through the slice builder, and produced coherent per-class Dice scores along with an overall lesion Dice of 0.8798. This fix was carried forward to a 100-epoch run (`yolo_2_5d_gpu_multiclass_fix_100ep`) and to a 2D-only multiclass run (`yolo_2d_multiclass_gpu_100ep`) to isolate the effect of through-plane context.

All YOLO Dice numbers reported in this thesis were computed by `eval_yolo_2_5d_dice.py`, which rasterizes predicted polygons per image and accumulates true-positive, false-positive, and false-negative pixel counts globally across all evaluated slices. The overall lesion Dice merges all non-background classes; per-class Dice is computed separately for class values 1, 2, and 3 (CET, NT, PE) and excludes background. This is a pixel-pooled, slice-level Dice and is therefore not directly comparable to the per-patient volumetric Dice produced by the SwinUNETR pipeline; the resulting comparability caveats are revisited in Sections 5 and 6.

4.5 Phase C: SwinUNETR Adaptation to Pretreat-MetsToBrain-Masks

Phase C adapted the volumetric SwinUNETR pipeline to the Pretreat-MetsToBrain-Masks

dataset through a sequence of principled design changes, each motivated by a concrete hypothesis about which factor was limiting performance. SwinUNETR is the primary volumetric architecture studied in this thesis, and the per-class results from this phase are reported as the main volumetric finding in Section 5.

The implementation used MONAI's reference SwinUNETR module, with the standard four-stage hierarchical Swin transformer encoder and a four-class output head corresponding to background, contrast-enhancing tumor, necrotic non-enhancing tumor core, and peritumoral edema[11]. Volumes were loaded directly from the original NIfTI files without resampling to a fixed spacing, intensity-normalized using MONAI's NormalizeIntensityd transform with channel-wise non-zero scaling, and randomly cropped to $96 \times 96 \times 96$ voxel patches during training. Data augmentation was limited to random spatial flips along all three axes (probability 0.5 per axis); no random intensity shifts or scales and no random noise were applied. Training used a hybrid freeze-then-unfreeze schedule: the SwinViT backbone was frozen for the first ten epochs while the segmentation head adapted, then unfrozen so that the full network could fine-tune jointly for the remaining epochs.

Three fine-tuning configurations were evaluated. The first, `swinunetr_pretreat_head_ft_v1`, fine-tuned on the single-channel T1c manifest for 100 epochs with the BraTS 2021 `fold1_model.pt` as initial weights, using one input channel. The second, `swinunetr_pretreat_head_ft_4ch_v1`, used the same architecture, optimizer settings, and training duration but switched to the four-channel manifest (T1, T1c, T2, FLAIR) with four input channels, also initialized from `fold1_model.pt`, to test whether richer input context would help with small and heterogeneous metastases. The third, `swinunetr_pretreat_head_ft_150ep_v2`, returned to the single-channel T1c manifest but extended training to 150 epochs and initialized from MONAI's self-supervised SwinViT weights (`model_swinvit.pt`, released alongside Tang et al.[21]), testing whether the gap between configurations 1 and 2 reflected under-training and a weaker initialization rather than an architectural ceiling.

4.6 Evaluation Protocol

Evaluation followed established conventions in the brain tumor segmentation literature, with metric definitions chosen to match what each model family naturally produces. For the SwinUNETR pipeline, the primary metric was the Dice similarity coefficient computed per region (tumor core, whole tumor, enhancing tumor) on the BraTS sanity check, and as the mean across the three foreground classes (CET, NT, PE; background excluded) on the Pretreat dataset. The single best validation Dice across all training epochs was reported as the result for each SwinUNETR configuration. For the YOLO pipeline, the primary metrics were the overall lesion Dice and per-class Dice for CET, NT, and PE, computed by the polygon-rasterization procedure described in Section 4.4; standard YOLO metrics (bounding-box and mask precision, recall, and mean average precision at IoU thresholds of 0.5 and 0.5–0.95) are also reported in Section 5.

Several caveats apply to direct comparison between the two pipelines and are flagged here so that the results in Section 5 can be read accurately. First, the SwinUNETR Dice is computed per patient on full 3D volumes and then averaged, while the YOLO Dice is a pixel-pooled global statistic over all slices; these are different quantities and small numerical differences should not be over-interpreted. Second, the SwinUNETR validation set was a 20-case patient-level holdout drawn from the 200-patient Pretreat dataset using a fixed random seed (`val_ratio` 0.1, seed 42), whereas the YOLO data partitioning operated at the slice level rather than the patient level, yielding a 196-case YOLO validation set whose patient identifiers overlap heavily with both the SwinUNETR training set and the YOLO training set. In practice, the YOLO validation set shares 176 patients with the SwinUNETR train-

ing set, so the cross-family comparison in Section 5 should be read as a comparison of two pipelines on overlapping but not identical Pretreat data, rather than as a head-to-head evaluation on a shared held-out cohort. The implications of this overlap are discussed further as a limitation in Section 6.5.

5 Results

This section presents the empirical findings from the three experimental phases described in Section 4. Phase A reports the smoke and sanity tests that verified the pipelines. Phase B reports the YOLO-based slice-level detection-and-segmentation results across single-class, 2D multiclass, and 2.5D multiclass formulations on Pretreat-MetsToBrain-Masks. Phase C reports the SwinUNETR volumetric segmentation results on the same dataset and is the main empirical contribution of the thesis. The section concludes with a side-by-side comparison and a discussion of cross-family metric comparability that follows on from the comparability caveats already introduced in Section 4.6.

5.1 Phase A: Smoke and Sanity Outcomes

The verification phase produced the expected qualitative pattern. The deliberately small smoke sanity run (`sanity_smoke_ssl`), which used SSL-pretrained SwinViT weights on a two-case BraTS slice, yielded a mean Dice of 0.0464. This score was expected to be low and is reported only to confirm that the SSL weights, the BraTS data loader, and the evaluation script all functioned end to end.

The full BraTS fold-1 sanity evaluation (`swinunetr_brats21_fold1_sanity`) provided the more informative check. Using the locally stored `fold1_model.pt` checkpoint on the 250-case fold-1 holdout, the SwinUNETR implementation achieved a Dice score of 0.9045 for tumor core, 0.9273 for whole tumor, and 0.8913 for enhancing tumor, for a mean Dice of 0.9077 across the three regions. These results are competitive with the published literature on BraTS and confirm that the implementation, training pipeline, and evaluation protocol are correct. Table 1 reports the per-region BraTS sanity results.

Table 1: Per-region BraTS 2021 fold-1 sanity-check Dice scores.

BraTS Region	Dice
Tumor Core (TC)	0.9045
Whole Tumor (WT)	0.9273
Enhancing Tumor (ET)	0.8913
Mean	0.9077

The YOLO smoke tests (`yolo_2_5d_smoketest` and `yolo_2_5d_smoketest2`) were one-epoch runs on a four-image dataset using the YOLOv8n-seg variant; both ran to completion (the second yielding an overall lesion Dice of 0.0000, as expected) and confirmed end-to-end pipeline functionality before the in-domain experiments.

5.2 Phase B: YOLO-Based Detection-and-Segmentation

The Phase B YOLO pipeline produced strong slice-based results, with an experimental progression that revealed how task formulation, dimensionality, and training duration each contributed to the final performance. The progression also exposed a class-formulation mismatch in an intermediate configuration that, once corrected, recovered the per-class signal expected from the dataset. All Dice numbers reported in this subsection are pixel-pooled, slice-level statistics computed by the polygon-rasterization procedure described in Section 4.4, and should be read with the comparability caveats given in Section 4.6 in mind.

Switching to the 2.5D multiclass formulation produced the largest single jump in the YOLO progression, although the change combined several factors at once: a corrected three-class label scheme, a 12-channel 2.5D input stack in place of single-channel 2D, and a switch to the `brainmets_2_5d.yaml` dataset definition.

The `yolo_2_5d_full` configuration (CPU, 50 epochs) raised the overall lesion Dice to 0.7304, with coherent contributions from all three foreground classes. The `yolo_2_5d_gpu` configuration extended this to GPU and reached an overall lesion Dice of 0.8761, but inspection of its per-class export revealed a systematic inconsistency in which the CET class scored low and the NT and PE classes scored zero, despite qualitatively reasonable overlays. As discussed in Section 4.4, this was traced to a class-formulation mismatch in which the underlying YOLO model had been trained with a binary (`nc=1`) head while the evaluation script expected the three-class Pretreat label scheme.

The fix was applied in `yolo_2_5d_gpu_multiclass_fix`, which enforced the three-class scheme in both training and evaluation and preserved multiclass labels through the slice builder. The corrected run reached an overall lesion Dice of 0.8798 with coherent per-class scores. Increasing this fixed configuration to 100 epochs (`yolo_2_5d_gpu_multiclass_fix_100ep`) further improved the result to an overall lesion Dice of 0.8956, the highest YOLO result indexed in this thesis, with per-class Dice of 0.861, 0.852, and 0.750 for CET, NT, and PE respectively. A 150-epoch extension (`yolo_2_5d_gpu_multiclass_fix_150ep`) was attempted but did not produce a completed metrics artifact and is therefore excluded from the quantitative comparison.

The 2D-only multiclass control (`yolo_2d_multiclass_gpu_100ep`) was run with the same 100-epoch GPU configuration but without the 2.5D through-plane context. It reached an overall lesion Dice of 0.8895 with per-class Dice of 0.858, 0.845, and 0.744, Figure 3 shows a representative output from this configuration. This is only marginally below the 2.5D result and indicates that, with sufficient training and a corrected class formulation, a pure 2D YOLO pipeline is competitive with the 2.5D variant in this setting; the through-plane context appears to provide a small but measurable benefit. Table 2 reports the YOLO results.

Three observations emerge from the YOLO progression. First, the largest gains came from corrections to task formulation rather than from architectural changes. Between `train2` (single-class label scheme, evaluated under a multiclass protocol, overall lesion Dice 0.2073) and `yolo_2_5d_gpu_multiclass_fix` (three-class label scheme, evaluation matched, overall lesion Dice 0.8798), several things changed at once: the class formulation moved from single-class to three-class, the input moved from 2D slices to a 2.5D stack, and the multiclass evaluation script was reconciled with the model's class head. The intermediate runs make clear that no single one of these changes can carry the whole jump, but together they recovered a coherent per-class signal that the earlier baselines simply could not produce. Second, with task formulation and training duration fixed, dimensionality alone (2D versus 2.5D) produces a small but measurable advantage in favor of 2.5D: the 100-epoch 2.5D run reached an overall lesion Dice of 0.8956 against 0.8895 for the 2D control, a difference of about 0.6 Dice points. This gap is consistent with the observation in the wider literature that limited through-plane context is helpful for segmenting small structures[8]. Third, training duration matters but with diminishing returns: the 100-epoch result improved on the 50-epoch result by about 1.6 Dice points (0.8956 against 0.8798), while the 150-epoch attempt did not yield a usable improvement within the time budget. The YOLO results form the slice-based comparator against which the volumetric SwinUNETR results in the next subsection are read.

5.3 Phase C: SwinUNETR Volumetric Segmentation (Main Result)

The SwinUNETR adaptation phase produced the main volumetric segmentation result of the thesis, and the progression of three runs makes the contribution of each design change reasonably easy to read. All Dice numbers in this subsection are per-patient volumetric statistics on the 20-case patient-level validation holdout (`val_ratio` 0.1, seed 42), measured as the mean

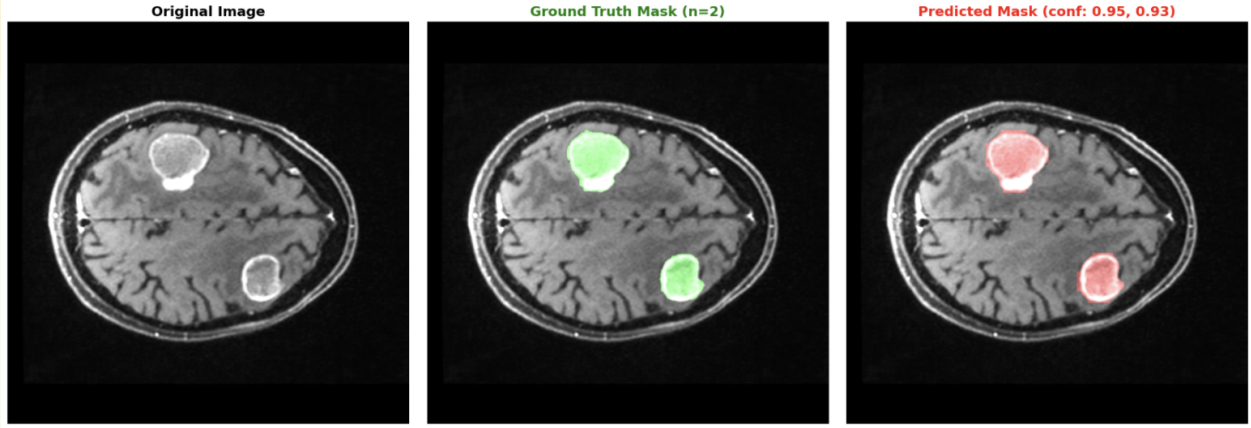

Figure 3: Qualitative example from the YOLOv8 2D model on the Pretreat-MetsToBrain-Masks dataset. Left: input T1c slice. Middle: ground-truth segmentation (green). Right: predicted segmentation (red) with high confidence scores (0.95, 0.93). The model accurately captures lesion location, shape, and extent across multiple metastases.

Table 2: YOLO experimental progression on Pretreat-MetsToBrain-Masks. Dice values are pixel-pooled slice-level statistics.

Configuration	Setup	Overall Dice	CET	NT	PE
`yolo_2_5d_full`	2.5D, CPU, 50ep	0.7304	–	–	–
`yolo_2_5d_gpu`	2.5D, GPU, 50ep (bug)	0.8761	low	0	0
`yolo_2_5d_gpu_multiclass_fix`	2.5D, GPU, 50ep, fixed	0.8798	0.854	0.835	0.727
`yolo_..._fix_100ep`	2.5D, GPU, 100ep, fixed	0.8956	0.861	0.852	0.750
`yolo_2d_multiclass_gpu_100ep`	2D, GPU, 100ep, fixed	0.8895	0.858	0.845	0.744

across the three foreground classes (CET, NT, PE; background excluded).

In the single-channel configuration using only the contrast-enhanced T1 sequence (`swinunetr_pretreat_head_ft_v1`, 100 epochs, `fold1_model.pt` initial weights), SwinUNETR achieved a best validation mean Dice of 0.6073 across the three foreground classes. This level of performance is consistent with the published observation that single-modality input is generally insufficient for high-quality multi-region brain tumor segmentation, since the contrast-enhanced T1 highlights enhancing tumor but provides limited information about non-enhancing components and edema[9,11].

In the four-channel configuration (`swinunetr_pretreat_head_ft_4ch_v1`, 100 epochs, `fold1_model.pt` initial weights), the same architecture trained for the same number of epochs on the same patient split achieved a best validation mean Dice of 0.7614, an improvement of more than fifteen Dice points over the single-channel result. Because non-enhancing core and peritumoral edema are most clearly visible on FLAIR and T2 rather than on contrast-enhanced T1, the addition of the three remaining sequences as input channels would be expected to help these classes in particular, and the magnitude of the overall gain is consistent with this interpretation. The result mirrors the findings of the wider brain tumor segmentation literature, which has consistently shown that multimodal input is essential for high-quality multi-region segmentation[8,11].

The third configuration (`swinunetr_pretreat_head_ft_150ep_v2`) returned to the single-channel T1c manifest but extended training to 150 epochs and initialized from MONAI's self-supervised SwinViT

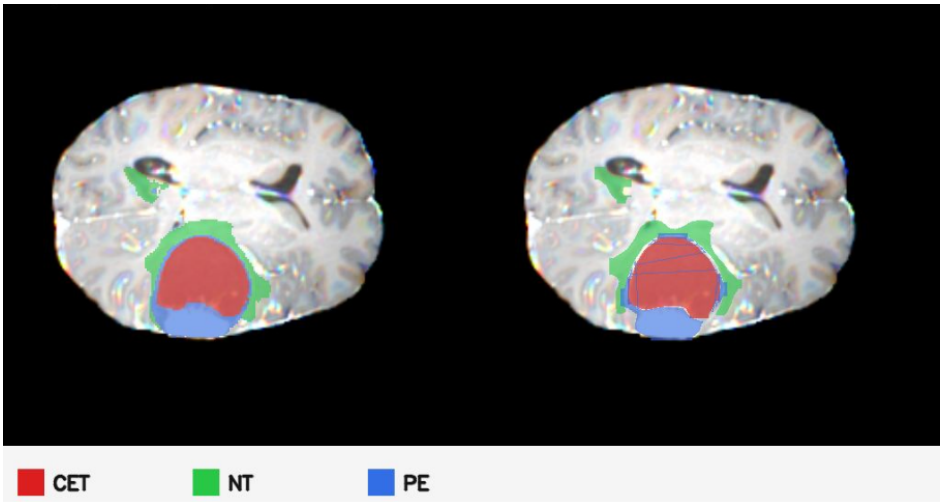

Figure 4: Qualitative multiclass segmentation from the 3D SwinUNETR model. Left: ground-truth segmentation. Right: model prediction (visualized as a representative axial slice). Red = contrast-enhancing tumor (CET), green = necrotic tumor core (NT), blue = peritumoral edema (PE). The model captures lesion location and subregion structure with strong spatial agreement.

weights (`model_swinvit.pt`, released alongside Tang et al.[21]) rather than the BraTS fold-1 checkpoint. This configuration achieved a best validation mean Dice of 0.7740, marginally exceeding the four-channel 100-epoch result and producing the highest in-domain Pretreat performance recorded in this thesis (Figure 4).

Because this run differed from the four-channel run on two axes simultaneously (training duration and pretraining initialization, but not input channels), the comparison is not a clean ablation and cannot be used to argue that single-channel SSL with 150 epochs is uniformly superior to four-channel 100 epochs. What it does support is the broader conclusion that adequate training duration and a strong initialization can recover some of the information that is missing from a single-channel input. A separate 150-epoch run without the SSL initialization (`swinunetr_pretreat_head_ft_150ep`) was also attempted but did not produce a final summary artifact and is excluded from the quantitative comparison. Table 3 summarizes the three SwinUNETR Pretreat configurations.

All three Phase C runs were trained on the same 180-case training subset and evaluated on the same 20-case validation subset, with the best validation Dice across all training epochs reported. The gap between the BraTS sanity-check result (mean Dice 0.9077) and the best Pretreat result (mean Dice 0.7740) is substantial and is consistent with two known phenomena in the brain tumor segmentation literature. First, BraTS has been heavily curated and uniformly preprocessed, and segmentation models tend to perform considerably better on BraTS than on real clinical data. Berkley and colleagues, working at Brown, reported a similar gap when applying a state-of-the-art 3D U-Net trained on BraTS to in-house clinical glioma cases, and observed that the lower clinical numbers were not statistically distinguishable from inter-rater variability between expert radiation oncologists[5]. Second, brain metastases are intrinsically more difficult to segment than gliomas: they are smaller, more numerous, more dispersed, and more heterogeneous in appearance, and per-lesion Dice on metastasis benchmarks is correspondingly lower[12].

5.4 Cross-Approach Comparison

Taken together, the three phases reveal a coherent pattern. Phase A confirmed that both pipelines functioned end to end before any modeling claim was made. Phase B showed

Table 3: SwinUNETR Pretreat fine-tuning configurations and best validation mean foreground Dice (CET, NT, PE; background excluded).

Configuration	Channels	Epochs	Init	Best Val Dice
T1c, 100ep	1 (T1c)	100	`fold1_model.pt`	0.6073
4-channel, 100ep	4 (all)	100	`fold1_model.pt`	0.7614
T1c, 150ep, SSL init	1 (T1c)	150	`model_swinvit.pt` (SSL)	0.7740

that the YOLO slice-based pipeline could be developed into a strong comparator through corrected multiclass formulation, 2.5D context, and longer training, reaching an overall lesion Dice of 0.8956 on the slice-level evaluation. Phase C showed that the SwinUNETR volumetric pipeline could be adapted to Pretreat-MetsToBrain-Masks with substantial gains from richer input channels (single-channel 0.6073 to four-channel 0.7614) and, separately, from longer training with self-supervised initialization (single-channel 100-epoch 0.6073 to single-channel 150-epoch SSL 0.7740), producing the main volumetric per-class segmentation result of the thesis.

The numerical comparison between SwinUNETR (best Pretreat mean foreground Dice 0.7740) and the YOLO 2.5D 100-epoch run (best Pretreat overall lesion Dice 0.8956) should not be read as a single-axis claim that YOLO outperforms SwinUNETR on this task, for two reasons developed in Section 4.6. First, the two metrics are computed at different spatial granularities: the SwinUNETR result is a per-patient volumetric mean across the three foreground classes on a 20-case patient-level hold-out, while the YOLO lesion Dice is a pixel-pooled global statistic over all evaluated slices in the YOLO validation set; small numerical differences between these quantities should not be over-interpreted. Second, the two pipelines were not evaluated on the same patient cohort: the YOLO data partitioning operated at the slice level, yielding a 196-case YOLO validation set whose patient identifiers overlap with the SwinUNETR training set on 176 cases. Reading the YOLO result as a head-to-head improvement over SwinUNETR would therefore confuse a difference in metric and split with a difference in modeling capability.

The two model families are best read as complementary views of the same underlying problem rather than as direct rivals. SwinUNETR produces volumetric, per-class consistent segmentations on full 3D volumes, which is the natural input for downstream volumetric measurements such as total tumor burden or per-region volume. YOLO produces fast, slice-wise predictions that are well suited to detection and triage workflows, where flagging the presence of metastases on individual slices is the relevant task. Both reached scientifically meaningful performance once their respective design issues were corrected, and the experimental progression tracked two distinct axes: increased modality coverage and longer training with a stronger initialization for SwinUNETR, and a corrected task formulation followed by added through-plane context for YOLO. The implications of this complementarity, including the comparability caveats and the split-overlap limitation, are explored in more detail in Section 6.

6 Discussion

This section steps back from the per-phase results and asks what the experiments collectively suggest about brain metastasis segmentation. Three methodological lessons emerge — the role of volumetric context, the value of multimodal input together with training duration and initialization, and the limits of foundation models in their default configurations — and these are followed by a discussion of clinical implications, the most consequential limitations of the present work, and natural future directions.

6.1 Why Volumetric Context Matters

A central methodological lesson of the thesis concerns the role of volumetric context. Two-dimensional models, even strong ones such as YOLOv8, see only a single axial slice at a time and must commit to a segmentation decision without access to information from neighboring slices. For brain metastases, this is a particularly costly limitation: many lesions are spherical or near-spherical and are visible across only two or three contiguous axial slices[12]. On the central slice, a lesion is at its largest and most distinctive, but on the bordering slices it is small and ambiguous, and a 2D model that segments each slice independently has no way to use the appearance of the central slice to inform its decisions on adjacent slices.

Three-dimensional architectures, in contrast, integrate information across the entire volume in a single forward pass. SwinUNETR's hierarchical encoder produces features at multiple spatial scales, with the deepest layers having receptive fields that span large portions of the volume. The resulting features can encode information such as whether a candidate enhancing focus is consistent with a metastasis (compact, roughly spherical, surrounded by a halo of edema) or with a vessel (linear, branching, no edema) without having to make this judgment on a single slice. The 2.5D YOLO results are interesting in this light: extending the input from a single slice to a stack of three adjacent T1c slices forming an RGB-style 3-channel input improved overall lesion Dice by about 0.6 points over the 2D control, a small but consistent gain. Limited through-plane context helps; full volumetric reasoning offers further potential, though the comparability caveats discussed in Sections 4.6 and 5 prevent direct numerical adjudication between the two.

6.2 The Value of Multimodal Input, Training Duration, and Initialization

The Phase C SwinUNETR experiments offer a relatively clean test of which factors most affect volumetric segmentation quality on Pretreat-MetsToBrain-Masks. Holding architecture, training data, and validation split constant, switching from a single-channel T1c input to a four-channel multimodal input raised the mean foreground Dice from 0.6073 to 0.7614, a fifteen-point improvement that exceeds the gap typically reported between architectural variants on brain metastasis benchmarks[12]. The size of the gain confirms that the four standard MRI sequences are not redundant: each contributes meaningful information that the model can use, and contrast-enhanced T1 alone is not sufficient for high-quality multi-region segmentation. This finding aligns with broader literature showing that multimodal models substantially outperform single-modality variants for brain tumor segmentation[8,11,12].

Extending training to 150 epochs and initializing from MONAI's self-supervised SwinViT weights (`model_swinvit.pt`, released alongside Tang et al.[21]) raised the best validation Dice to 0.7740 in a separate single-channel 150-epoch SSL configuration, marginally exceeding the four-channel 100-epoch result. As noted in Section 5.3, this comparison is not a clean ablation because two factors changed simultaneously (training duration and pretraining initialization), so the marginal improvement cannot be cleanly attributed to either alone. What it does suggest is that adequate training duration combined with a strong initialization can recover some of the information that is missing from a single-channel input. The clinical implication is straightforward: when deployment-time data permit, automated brain metastasis segmentation systems should ingest all four standard MRI sequences rather than the contrast-enhanced T1 alone, and when modality access is limited, longer training with self-supervised pretraining is a partial substitute for the missing channels.

6.3 Limitations of Foundation Models in this Setting

Recent work on foundation models for medical image segmentation has raised the possibility that large pretrained models, used either zero-shot or with minimal fine-tuning, could substitute for task-specific supervised training. Reviews of this question on brain tumor data have been mixed, with strong performance reported when prompts are provided or domain-specific fine-tuning is applied, and weaker performance reported in pure zero-shot settings[8]. Preliminary exploration in this thesis with SAM2 and the original Segment Anything Model was consistent with that pattern: without task-specific prompts, foundation models trained on natural images tend to segment the most salient structure in the field of view, which on an axial brain MRI is the brain itself rather than the small enhancing foci of interest. Brain metastases are not salient in the natural-image sense — they are small, often blend into surrounding tissue, and require an understanding of anatomy and pathology that is not present in the training distribution of these models. This observation motivated the focus on supervised volumetric and slice-based pipelines reported in this thesis.

This does not mean foundation models are useless for brain metastasis segmentation. The literature shows that with appropriate prompts, fine-tuning, or domain-specific pretraining, models in the SAM family can be competitive with supervised baselines on a variety of medical tasks[8]. The limitation observed in the preliminary work here is specific to the prompt-free configuration, and a natural follow-up, discussed in Section 6.5, is to combine the YOLOv8 detector, which produces high-quality bounding-box predictions for individual lesions, with a prompted foundation-model segmenter such as SAM2.

6.4 Practical Implications for Clinical Translation

Reading these results through a clinical lens, several practical points emerge. The best Pretreat result of mean Dice 0.7740 is encouraging for an early-stage volumetric pipeline on brain metastases, but it is not yet at a level where autonomous clinical use would be appropriate. Most of the experiments in this thesis were performed on a single dataset from a single institution, and as the wider literature has emphasized, multi-center validation is essential before any segmentation system can be considered for deployment[5,12].

A second point concerns the gap between curated benchmarks and clinical data. The BraTS sanity-check result (mean Dice 0.9077) and the best Pretreat result (mean Dice 0.7740) sit roughly thirteen Dice points apart. Some of this gap reflects the well-documented fact that BraTS has been heavily curated and uniformly preprocessed and that segmentation models tend to perform considerably better on curated benchmarks than on real clinical data, which Berkley and colleagues showed directly when applying a state-of-the-art 3D U-Net trained on BraTS to in-house clinical glioma cases[5]. The remainder of the gap reflects the intrinsic difficulty of brain metastasis segmentation relative to glioma segmentation: metastases are smaller, more numerous, more dispersed, and more heterogeneous in appearance, and per-lesion Dice on metastasis benchmarks is correspondingly lower[12]. Reporting both numbers together, rather than only the higher one, gives a more honest picture of where this kind of pipeline currently stands.

A third practical point concerns the role of slice-based pipelines. The YOLO 2.5D multiclass results (overall lesion Dice 0.8956 on the slice-pooled metric) suggest that slice-based detection-and-segmentation systems remain useful as fast, lightweight components within a larger workflow, particularly for detection and triage tasks where flagging the presence of metastases on individual slices is the operational requirement. Volumetric pipelines such as SwinUNETR and slice-based pipelines such as YOLO are best read as complementary rather than competing: the former is the natural input for downstream volumetric measurements such as total tumor burden or per-region

volume, while the latter is well suited to high-throughput screening and review.

6.5 Limitations and Future Work

Several limitations of the present thesis should be acknowledged, and most of them suggest natural directions for follow-up work. The most consequential is the way in which the SwinUNETR and YOLO experiments were partitioned. As described in Section 4.6, the SwinUNETR validation set was a 20-case patient-level holdout drawn with a fixed random seed, whereas the YOLO data partitioning operated at the slice level rather than the patient level. The result is a 196-case YOLO validation set that shares 176 patients with the SwinUNETR training set. The cross-family comparison in Section 5.4 should therefore be read as a comparison of two pipelines on overlapping but not identical Pretreat data, rather than as a head-to-head evaluation on a shared held-out cohort. A natural follow-up would be to re-run the YOLO experiments under a strict patient-level split that matches the SwinUNETR holdout exactly, which would put the cross-family comparison on cleaner footing. An intermediate step would be to report YOLO results on the 20-patient subset of the YOLO validation set that is not in the SwinUNETR training set, providing a smaller but methodologically cleaner cross-family comparison.

A second limitation is that the SwinUNETR Phase C runs explored two different paths from the single-channel 100-epoch baseline — one varying input channels (single-channel to four-channel), the other varying training duration and pretraining initialization (100-epoch `fold1_model.pt` initialization to 150-epoch SSL initialization) — but did not isolate training duration or initialization as independent factors. A targeted ablation that varies training duration alone, with channel count and initialization held fixed, would tighten the architectural argument. A third limitation is that the YOLO Phase B results, although improved substantially after the multi-class class-formulation fix, were exposed to several sources of methodological variability earlier in the progression — an unusable single-class CPU baseline and the `nc=1` versus three-class evaluation mismatch in `yolo_2_5d_gpu` — that were corrected only after they were observed. A more disciplined experimental design would have evaluated the corrected configuration first, with the earlier baselines reported only for completeness.

A fourth limitation is the absence of an explicit comparison against a 3D U-Net baseline. The decision to focus on SwinUNETR was motivated by its strong reported performance on BraTS-style tasks[11], but a head-to-head comparison with a well-tuned 3D U-Net or nnU-Net on Pretreat-MetsToBrain-Masks would have made the architectural argument tighter. nnU-Net is the natural candidate because of its automated configuration and its track record on brain tumor benchmarks[19]. Adding such a baseline is a clear next step and would also help disentangle whether the gains observed in this thesis are specific to the transformer-based encoder or would have appeared with any sufficiently well-trained 3D segmentation model.

A natural extension at the dataset level is to bring the Yale-Brain-Mets-Longitudinal collection back into the experimental loop. The 898-case post-treatment subset prepared during this project provides infrastructure for deployment-style inference on a much larger and more heterogeneous cohort than Pretreat. The cohort lacks ground-truth segmentations, so the most immediate priority would be to obtain partial annotations on a tractable subset, either through expert review of model overlays or through targeted annotation by collaborating clinicians, and to compute proper Dice and lesion-wise metrics on that subset. Beyond simple inference, the longitudinal nature of the dataset makes it especially well suited for methods that exploit prior imaging. Approaches such as DeepMedic+ from Huang and colleagues, in which the segmentation mask from a prior visit is provided to the network as an additional input channel, have been shown to substantially improve detection of small le-

sions on follow-up MRI[23]. Implementing a longitudinal variant of the SwinUNETR pipeline that takes a prior segmentation as an additional channel and outputs an updated segmentation for the current visit would be a natural extension, and the SPIRS architecture from Patel and colleagues, which jointly performs registration and segmentation, points to an even more sophisticated way of leveraging prior imaging[24].

Two further directions follow on from the foundation-model discussion in Section 6.3. First, building on the brief SAM2 exploration discussed there, the combination of a YOLOv8 detector that locates lesions accurately at the slice level with a prompted foundation-model segmenter that produces precise boundaries within each box is a clear opportunity for future work. The detector supplies boxes; the prompted segmenter refines masks within those boxes; and the resulting two-stage pipeline could potentially deliver both the recall of supervised detection and the precise boundaries of the foundation model. The literature has already begun to explore similar pipelines in other medical-imaging domains[8], and a brain-metastasis-specific version is a clear opportunity. Second, although the choice in this thesis was between strict 2D and strict 3D, a more systematic exploration of intermediate dimensionalities is worth pursuing: 2.5D with varying numbers of adjacent slices, anisotropic 3D patches, and slice-thickness-aware sampling all sit between the two extremes studied here. Reviews of 2.5D versus 3D segmentation suggest that the optimal choice depends on dataset size, slice anisotropy, and the size of the target structures[8], and the heavily anisotropic Yale-BrainMets data in particular would be a useful benchmark for this question.

Finally, broader generalization testing remains essential. The Berkley and colleagues study showed that even state-of-the-art models trained on BraTS can lose substantial performance when evaluated on out-of-distribution clinical data, although they also found that the loss was no greater than inter-rater variability between expert clinical radiation oncologists[5]. Validating the four-channel SwinUNETR pipeline on additional public brain metastasis datasets, such as UCSF-BMSR or the BraTS-METS 2023 and 2025 collections[6,10], would test the robustness of the conclusions reported here and would put the work into direct dialogue with the most recent benchmarks in the field.

References

[1] Q. T. Ostrom, C. H. Wright, and J. S. Barnholtz-Sloan. Brain metastases: epidemiology. *Handbook of Clinical Neurology*, 149:27–42, 2018. doi: 10.1016/B978-0-12-811161-1.00002-5.

[2] A. S. Achrol, R. C. Rennert, C. Anders, R. Soffietti, M. S. Ahluwalia, L. Nayak, S. Peters, N. D. Arvold, G. R. Harsh, P. S. Steeg, and S. D. Chang. Brain metastases. *Nature Reviews Disease Primers*, 5(1):5, 2019. doi: 10.1038/s41572-018-0055-y.

[3] W. B. Pope. Brain metastases: neuroimaging. *Handbook of Clinical Neurology*, 149:89–112, 2018. doi: 10.1016/B978-0-12-811161-1.00007-4.

[4] M. A. Vogelbaum, P. D. Brown, H. Messersmith, P. K. Brastianos, S. Burri, D. Cahill, I. F. Dunn, L. E. Gaspar, N. T. N. Gatson, V. Gondi, J. T. Jordan, A. B. Lassman, J. Maues, N. Mohile, N. Redjal, G. Stevens, E. Sulman, M. van den Bent, H. J. Wallace, J. S. Weinberg, G. Zadeh, and D. Schiff. Treatment for brain metastases: ASCO-SNO-ASTRO guideline. *Journal of Clinical Oncology*, 40(5):492–516, 2022. doi: 10.1200/JCO.21.02314.

[5] A. Berkley, C. Saueressig, U. Shukla, I. Chowdhury, A. Munoz-Gauna, O. Shehu, R. Singh, and R. Munbodh. Clinical capability of modern brain tumor segmentation models. *Medical Physics*, 50(8):4943–4959, 2023. doi: 10.1002/mp.16321.

[6] A. W. Moawad, A. Janas, U. Baid, D. Ramakrishnan, R. Saluja, N. Ashraf, N. Maleki, L. Jekel, N. Yordanov, P. Fehringer, A. Gkampenis, et al. The brain tumor segmentation – metastases (BraTS-METS) challenge 2023: Brain metastasis segmentation on pre-treatment MRI. *arXiv preprint arXiv:2306.00838*, 2024. URL https://arxiv.org/abs/2306.00838. Version 3.

[7] X. Luo, Y. Yang, S. Yin, H. Li, Y. Shao, D. Zheng, X. Li, J. Li, W. Fan, J. Li, X. Ban, S. Lian, Y. Zhang, Q. Yang, W. Zhang, C. Zhang, L. Ma, Y. Luo, F. Zhou, S. Wang, C. Lin, J. Li, M. Luo, J. He, G. Xu, Y. Gao, D. Shen, Y. Sun, Y. Mou, R. Zhang, and C. Xie. Automated segmentation of brain metastases with deep learning: A multi-center, randomized crossover, multi-reader evaluation study. *Neuro-Oncology*, 26(11):2140–2151, 2024. doi: 10.1093/neuonc/noae113.

[8] F. J. Dorfner, J. B. Patel, J. Kalpathy-Cramer, E. R. Gerstner, and C. P. Bridge. A review of deep learning for brain tumor analysis in MRI. *npj Precision Oncology*, 9(1):2, 2025. doi: 10.1038/s41698-024-00789-2.

[9] T. Yang et al. Progress and trends in neurological disorders research based on deep learning. *Computerized Medical Imaging and Graphics*, 116:102400, 2024. doi: 10.1016/j.compmedimag.2024.102400.

[10] N. Maleki, R. Amiruddin, A. W. Moawad, N. Yordanov, A. Gkampenis, P. Fehringer, F. Umeh, C. Chukwurah, F. Memon, B. Petrovic, B. Albalooshy, J. Cramer, M. Krycia, E. B. Shrickel, I. Ikuta, G. Thompson, L. Vidal, V. Kosovic, A. E. Goldman-Yassen, V. Hill, et al. Analysis of the MICCAI brain tumor segmentation – metastases (BraTS-METS) 2025 lighthouse challenge: Brain metastasis segmentation on pre- and post-treatment MRI. *arXiv preprint arXiv:2504.12527*, 2025. URL https://arxiv.org/abs/2504.12527.

[11] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2021)*,

volume 12962 of *Lecture Notes in Computer Science*, pages 272–284. Springer, 2022. doi: 10.1007/978-3-031-08999-2_22.

[12] T.-W. Wang, M.-S. Hsu, W.-K. Lee, H.-C. Pan, H.-C. Yang, C.-C. Lee, and Y.-T. Wu. Brain metastasis tumor segmentation and detection using deep learning algorithms: A systematic review and meta-analysis. *Radiotherapy and Oncology*, 190:110007, 2024. doi: 10.1016/j.radonc.2023.110007.

[13] D. Ramakrishnan, L. Jekel, S. Chadha, A. Janas, H. Moy, N. Maleki, M. Sala, M. Kaur, G. Cassinelli Petersen, S. Merkaj, M. von Reppert, C. Kirsch, M. Davis, M. Lin, F. Memon, and M. S. Aboian. A large open access dataset of brain metastasis 3D segmentations on MRI with clinical and imaging information. *Scientific Data*, 11:254, 2024. doi: 10.1038/s41597-024-03021-9.

[14] P. D. Brown, V. Gondi, S. Pugh, W. A. Tome, J. S. Wefel, T. S. Armstrong, J. A. Bovi, C. Robinson, A. Konski, D. Khuntia, D. Grosshans, T. L. S. Benzinger, D. Bruner, M. R. Gilbert, D. Roberge, V. Kundapur, K. Devisetty, S. Shah, K. Usuki, B. M. Anderson, B. Stea, H. Yoon, J. Li, N. N. Laack, T. J. Kruser, S. J. Chmura, W. Shi, S. Deshmukh, M. P. Mehta, and L. A. Kachnic. Hippocampal avoidance during whole-brain radiotherapy plus memantine for patients with brain metastases: Phase III trial NRG oncology CC001. *Journal of Clinical Oncology*, 38(10):1019–1029, 2020. doi: 10.1200/JCO.19.02767.

[15] M. Aboian, K. Bousabarah, E. Kazarian, T. Zeevi, W. Holler, S. Merkaj, G. Cassinelli Petersen, R. Bahar, H. Subramanian, P. Sunku, E. Schrickel, J. Bhawnani, M. Zawalich, A. Mahajan, A. Malhotra, S. Payabvash, I. Ikuta, M. Buono, M. von Reppert, D. Ramakrishnan, et al. Yale-brain-mets-longitudinal: A longitudinal mri dataset of brain metastases. *arXiv preprint arXiv:2506.14021*, 2025. URL https://arxiv.org/abs/2506.14021.

[16] N. A. Wijetunga and T. J. Yang. A radiation oncology approach to brain metastases. *Frontiers in Neurology*, 11:801, 2020. doi: 10.3389/fneur.2020.00801.

[17] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

[18] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. UNETR: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 574–584, 2022. doi: 10.1109/WACV51458.2022.00181.

[19] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. doi: 10.1038/s41592-020-01008-z.

[20] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li. TransBTS: Multimodal brain tumor segmentation using transformer. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, volume 12901 of *Lecture Notes in Computer Science*, pages 109–119. Springer, 2021. doi: 10.1007/978-3-030-87193-2_11.

[21] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh. Self-supervised pre-training of swin transformers for 3D medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20730–20740, 2022.

[22] H. Yu, Z. Zhang, W. Xia, Y. Liu, L. Liu, W. Luo, J. Zhou, and Y. Zhang. DeSeg: auto detector-based segmentation for brain metastases. *Physics in Medicine & Biology*, 68(2):025001, 2023. doi: 10.1088/1361-6560/acace7.

[23] Y. Huang, C. Bert, P. Sommer, B. Frey, U. Gaipl, L. V. Distel, T. Weissmann, M. Uder, M. A. Schmidt, A. Doerfler, A. Maier, R. Fietkau, and F. Putz. Deep learning for brain metastasis detection and segmentation in longitudinal MRI data. *Medical Physics*, 49(9):5773–5786, 2022. doi: 10.1002/mp.15863.

[24] J. Patel, S. R. Ahmed, K. Chang, P. Singh, M. Gidwani, K. Hoebel, A. Kim, C. Bridge, C. Teng, X. Li, G. Xu, M. McDonald, A. Aizer, W. L. Bi, I. Ly, B. Rosen, P. Brastianos, R. Huang, E. Gerstner, and J. Kalpathy-Cramer. A deep learning based framework for joint image registration and segmentation of brain metastases on magnetic resonance imaging (SPIRS). In *Proceedings of the 8th Machine Learning for Healthcare Conference*, volume 219 of *Proceedings of Machine Learning Research*, pages 565–587, 2023. URL https://proceedings.mlr.press/v219/patel23a.html.

[25] C. Saueressig, A. Berkley, R. Munbodh, and R. Singh. A joint graph and image convolution network for automatic brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2021)*, volume 12962 of *Lecture Notes in Computer Science*, pages 356–365. Springer, 2022. doi: 10.1007/978-3-031-08999-2_30.