ALMA MATER STUDIORUM

UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF COMPUTER SCIENCE

AND ENGINEERING

ARTIFICIAL INTELLIGENCE

MASTER THESIS

in

Big Data Analytics And Text Mining

MULTIMODAL DEEP LEARNING FOR MEDICAL IMAGING: A SURVEY AND A NEW APPROACH TO BRAIN TUMOR SEGMENTATION WITH INCOMPLETE DATA

CANDIDATE	SUPERVISOR
Enze Ge	Prof. Stefano Lodi

Academic year 2024-2025

Session 1st

Abstract

Multimodal MRI is crucial for brain tumor segmentation, but its clinical use is hampered by the ”missing modality problem,” where incomplete data degrades model performance and deployment. This thesis introduces the Grouped Modality Distillation Transformer (GMD-Trans), a novel, fully supervised framework designed to be inherently robust to this challenge. The GMD-Trans architecture uses a 3D Vision Transformer (ViT) backbone for global context modeling and a dual-stream encoder for synergistic modality groups. Features are integrated via a cross-attention mixer (IG-CAM). Robustness is achieved through a teacher-student knowledge distillation (KD) scheme guided by the mathematically stable Hölder Divergence to ensure performance even when key modalities are absent. Evaluated on the BraTS 2021 benchmark with randomly missing modalities, GMD-Trans achieves a state-of-the-art Dice score of 82.1% for the Tumor Core (TC), surpassing strong baselines. Ablation studies confirm the efficacy of the proposed methods. This specialized success, however, reveals a performance trade-off, with lower accuracy on the Enhancing Tumor (ET) region. GMD-Trans provides a powerful and efficient solution for robustly segmenting the main tumor body from incomplete data using a fully supervised paradigm, without needing complex pre-training or data synthesis. This work advances the development of dependable AI tools for real-world neuro-oncology.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Context

The field of multi-modal learning represents a significant and rapidly evolving frontier in artificial intelligence, with the primary objective of endowing machines with a more comprehensive and nuanced comprehension of their environment, mirroring the integrated nature of human perception. Its core principle involves the synthesis and analysis of information from disparate and heterogeneous sources, including but not limited to visual, textual, and acoustic data streams [1] [2]. The central impetus for this approach stems from the recognition that any single source of information, when considered in isolation, is often insufficient or subject to ambiguity. By strategically fusing these complementary data streams, multi-modal systems can overcome the limitations of individual modalities, leading to the creation of richer, more reliable, and more complete internal representations of complex real-world phenomena [3]. Achieving this requires a sophisticated, interdisciplinary approach that draws upon expertise from computer vision, natural language processing, and speech analysis, among other domains. The overarching ambition is the development of artificial intelligence systems capable of perception, reasoning, and interaction that more closely emulate the fluid and context-aware nature of human cognition.

With the maturation of this research area, the need for clear and rigorous foundational concepts has become paramount. A central element requiring formalization is the definition of ”modality” itself. Beyond the intuitive notion of simply involving multiple data formats, a more precise, functional definition is essential for methodical progress. Contemporary research suggests that modalities should be defined not merely by their format but by the distinct representational value they offer for a given task [1]. For instance, while the modalities for a visual question answering system are the image and its corresponding text, an autonomous vehicle might rely on a combination of visual camera data, LiDAR point clouds, and radar signals. This task-oriented perspective is vital for designing effective and targeted model architectures. Of equal importance is the establishment of systematic taxonomies for classifying the field's wide array of methodologies [2]. Such frameworks are indispensable for comparing different approaches, identifying research gaps, and understanding the conceptual relationships between various models. These classifications are frequently organized around critical architectural decisions, most notably the stage at which data fusion occurs—such as early, intermediate, or late fusion —as well as the methods used for learning coordinated representations and the specific learning objectives like alignment or co-learning.

Figure 1.1 illustrates the core challenges of multimodal learning. It begins with diverse input modalities (visual, text, audio, sensor) that are processed through representation and alignment. These foundational steps then support subsequent complex tasks including reasoning, generation, and transference, while quantification provides analytical evaluation across the framework.

Deep learning continuously plays a major role in multi-modal systems. It brings strong architectures that can translate one modality into another, improve representation learning, and handle multiple modalities at once [3]. These methods have been used in many fields, such as healthcare, robotics, video understanding, and text-to-image generation. All of these show that multi-modal learning has great potential in real applications.

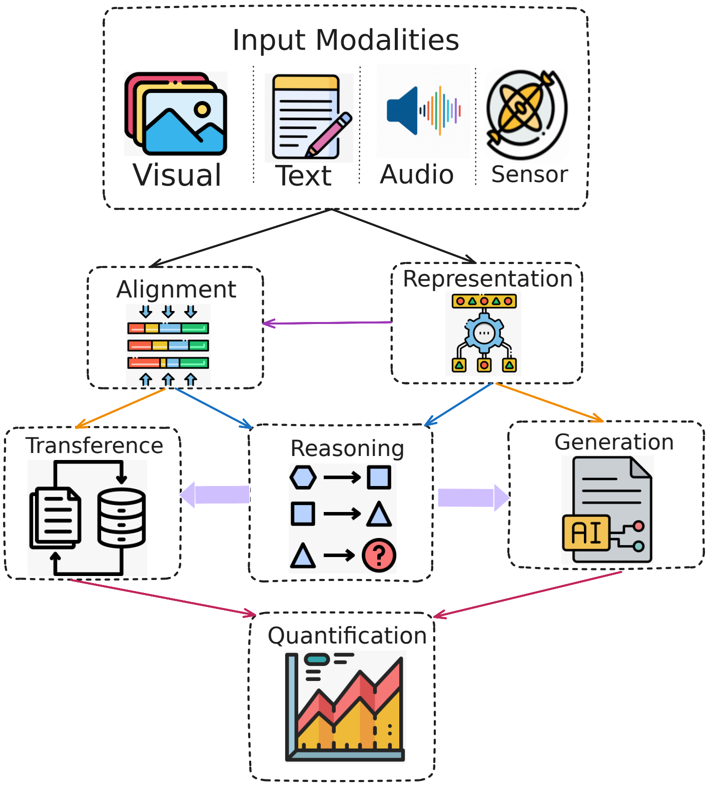


Figure 1.1: The core technical challenges in multi-modal learning.

Medical image segmentation represents a fundamental task in computer-aided diagnosis, enabling precise identification and delineation of anatomical structures and pathological regions. The field has experienced transformative advancement through deep learning technologies, particularly in addressing complex multimodal scenarios where information from different imaging modalities must be integrated effectively[4].

While conventional segmentation algorithms offer computational efficiency and interpretability, they often prove inadequate when faced with the inherent complexity, noise, and variability of medical imagery. The advent of deep learning has precipitated a paradigm shift in this domain, facilitating the development of segmentation solutions with substantially greater robustness and accuracy across a wide range of clinical applications [5].

Despite these advancements, several critical challenges persist as active areas of research. A primary unresolved issue is the design of more generalized and powerful models capable of robustly handling scenarios where certain modalities are unavailable during inference [6]. Concurrently, developing efficient and effective methods to align, integrate, and reason across disparate data sources remains a significant technical hurdle [7]. Moreover, for these sophisticated models to be viable for large-scale clinical deployment, they must satisfy stringent requirements for both scalability and computational efficiency.

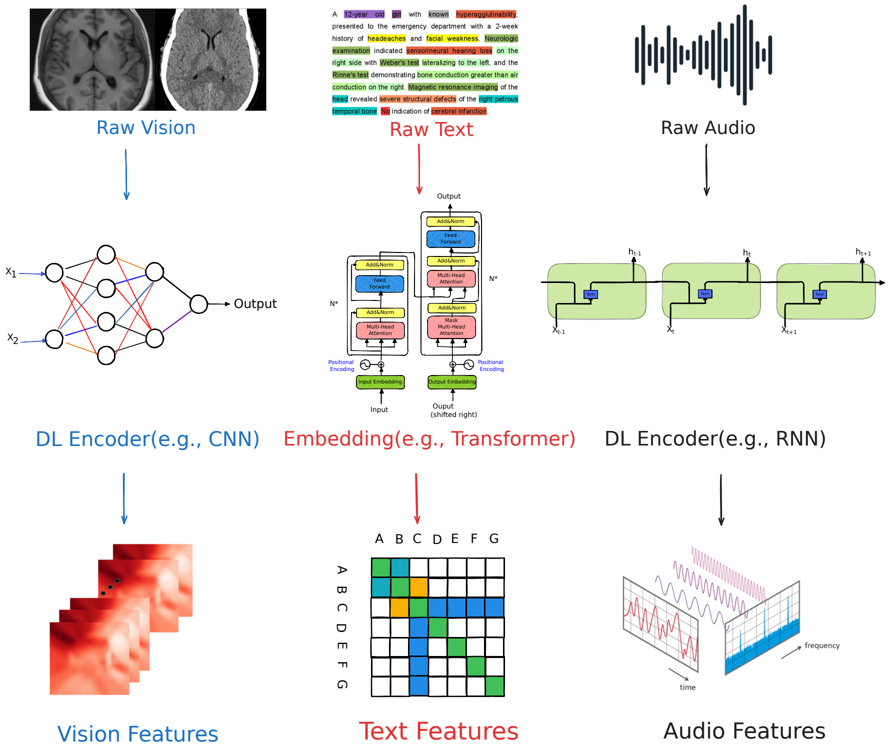
1.2 Foundations of Multi-modal Processing

The foundational principles of multi-modal processing are centered on the effective integration and interpretation of information from heterogeneous sources, such as text, images, audio, and video streams. The primary goal is to build computational models that can process and relate information from these diverse modalities to achieve a more robust and holistic understanding of complex phenomena. Recent advancements have led to significant progress, but the field is defined by a set of core technical challenges that must be addressed to build effective systems. A prominent taxonomy identifies six such challenges: representation, alignment, reasoning, generation, transference, and quantification [7], underscoring the breadth and complexity of the domain.

Representation learning is a cornerstone of multimodal processing, as it addresses the primary challenge of transforming heterogeneous data into a format that can be jointly processed. The objective is to learn representations that can capture and exploit the complementarity and redundancy of information across different modalities [1]. These representations can be broadly categorized into two main types:

Joint Representations: These methods project data from multiple modalities into a shared semantic space. This creates a unified representation where information from different sources can be directly compared and combined.

Coordinated Representations: In contrast to forcing all modalities into a single space, this approach learns separate representations for each modality while enforcing constraints to keep them correlated.



A 12-year old girl with known hyperagglutinability, presented to the emergency department with a 2-week history of headeaches and facial weakness. Neurologic examination indicated sensorineural hearing loss on the right side with Weber's test lateralizing to the left, and the Rinne's test demonstrating bone conduction greater than air conduction on the right. Magnetic resonance imaging of the head revealed severe structural defects of the right petrous temporal bone. No indication of cerebral infarction.

Figure 1.2: Unimodal representation learning.

Figure 1.2 shows the initial stage of unimodal representation learning, where raw Visual, Text, and Audio inputs are processed by distinct deep learning encoders. Modality-specific architectures transform these diverse inputs into structured Vision Features, Text Features, and Audio Features for subsequent multi-modal integration.

One of the most significant conceptual advancements is the adoption of task-relative definitions of multimodality, which prioritize the functional value of information for a specific objective over its format [1]. In medical imaging, for instance, this perspective is critical; for brain tumor segmentation, T1-weighted and FLAIR MRI scans are considered distinct modalities not merely because they are different imaging sequences, but because they provide unique, non-redundant information essential for delineating the tumor core and surrounding edema, respectively. This task-centric view promotes more flexible and effective system designs. Furthermore, surveys on multi-modal co-learning are systematically addressing persistent challenges like incomplete or noisy data, offering new taxonomies of techniques and applications that help to structure and guide future research directions [8].

A parallel and increasingly vital line of inquiry is the pursuit of model robustness. Dedicated frameworks are being developed to analyze and improve the stability of multi-modal systems under diverse and non-ideal conditions, which is crucial for their deployment in safety-critical applications like medicine [6]. At the same time, the potential of multi-modal large language models (MLLMs) is attracting considerable attention, particularly for tasks such as generating diagnostic reports from radiological images, which requires a deep fusion of visual and linguistic understanding [9]. The long-term ambition is to create generalist multi-modal AI systems that can adeptly handle a wide variety of tasks and modalities, representing an exciting and promising frontier for the field [10].

However, significant challenges and limitations remain. Scalability continues to be a major concern as models grow in complexity. The problem of missing or noisy modalities is a persistent practical barrier, especially in clinical workflows where a full complement of data is not always available for every patient. Furthermore, there is still a lack of universally accepted standards for evaluating multi-modal models, making direct comparisons between different approaches difficult, [11]. To address these issues, researchers are actively exploring methods such as correlation maximization or minimization to refine feature representations [11], as well as developing improved data processing techniques specifically designed for modern multi-modal architectures [12].

In conclusion, the foundations of multi-modal processing are built upon a rich confluence of concepts, techniques, and applications. With sustained research, significant improvements in how cross-modal data is understood and integrated within higher-level models are anticipated [13], [9]. These advances will undoubtedly influence numerous domains, including healthcare, education, and entertainment, underscoring the importance of continued support and funding for this pivotal field.

Chapter 2

Deep Learning for Multi-modal Data

The emergence of deep learning has fundamentally transformed the processing of multimodal data, enabling sophisticated analysis of heterogeneous data sources across diverse domains [3, 14]. This paradigm has proven particularly transformative in medical imaging, where clinicians routinely analyze multiple imaging sequences to achieve comprehensive tumor characterization.The integration of deep learning with medical imaging has initiated a paradigm shift in computational medicine, moving beyond simple image reconstruction to sophisticated analysis and diagnostic support. This evolution is particularly pronounced in the domain of multimodal imaging, where the synthesis of information from disparate sources offers a more holistic view of complex pathologies than any single modality can provide alone. This section delineates the clinical rationale for multimodal analysis, outlines the foundational challenges inherent to medical data that shape algorithmic development, and introduces the key technological paradigms that have emerged to address these hurdles.

Modern multimodal deep learning systems must navigate several core technical challenges: representation learning for heterogeneous data types, alignment of information across modalities, and the generation of cross-modal content. These challenges become particularly pronounced in medical applications where data heterogeneity, missing modalities, and stringent accuracy requirements create unique computational demands.

2.1 The Clinical Imperative for Multimodality: Beyond Single-View Diagnostics

Medical imaging is a cornerstone of modern clinical practice, offering non-invasive windows into the human body's internal structures and biological processes. However, the diagnostic power of any single imaging modality is inherently constrained by its underlying physical principles. Consequently, a single modality often fails to present a complete characterization of a specific organ or lesion, thereby limiting the accuracy and comprehensiveness of clinical diagnosis.

Multimodal medical image analysis addresses this fundamental limitation by integrating complementary information from different imaging sources. This approach mirrors the diagnostic process of a clinician, who synthesizes data from a multitude of resources—including radiological images, histopathology slides, electronic health records, and genomic data—to arrive at a treatment decision[15]. The objective is to transform the imaging paradigm from a conventional ”what you see is what you get” model to a more adaptive and clinically potent ”what you see is what you need” approach. By fusing data, it becomes possible to visualize previously unobservable targets, overcome the physical limitations of individual hardware systems, and enable advanced visual tasks such as precise semantic segmentation and three-dimensional reconstruction.The growing consensus is that future computer-assisted diagnostic systems must be capable of processing multimodal data simultaneously to achieve their full potential[15].

2.2 Architectural Foundations and Evolution

2.2.1 Convolutional Neural Network Advances

Traditional convolutional neural networks, particularly U-Net[16] and its variants, established the foundation for medical image segmentation through their encoder-decoder architecture with skip connections. These architectures have demonstrated remarkable effectiveness in preserving spatial information while capturing hierarchical features. Recent advances have introduced sophisticated modifications like the Modified Connected U-Net with Guided Decoder (MCU-Net-GD)[17], which utilizes a dual U-Net structure with guided decoder mechanisms to enhance segmentation performance.

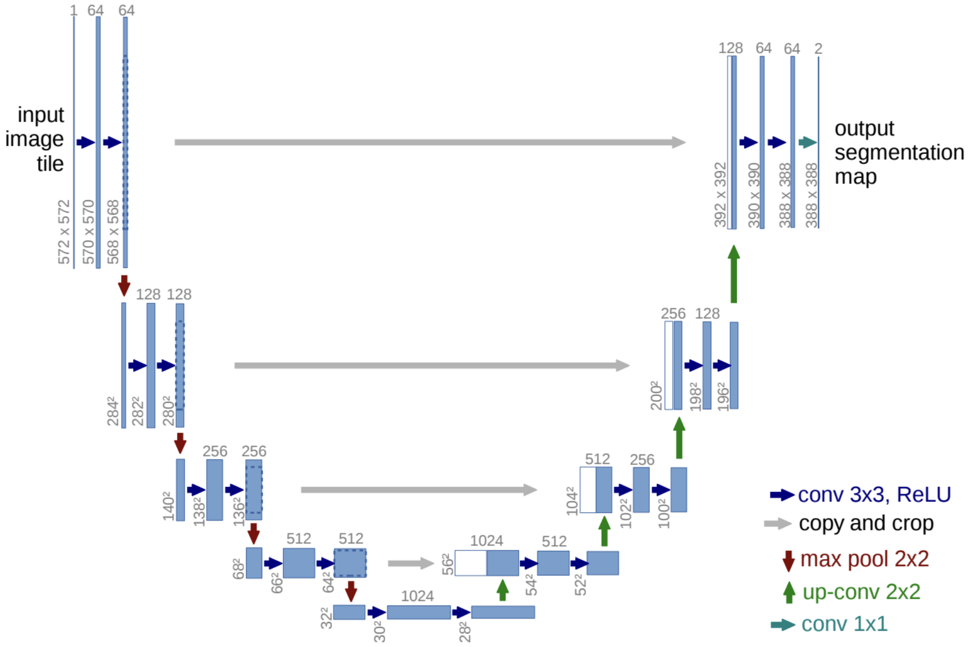


Figure 2.1: Classic U-Net model architecture[16].

However, CNNs face inherent limitations in modeling long-range dependencies due to the locality of convolution operations. This constraint has motivated the development of hybrid architectures. Advanced CNN approaches have also incorporated Atrous Spatial Pyramid Pooling (ASPP)[18] blocks to gather contextual information across multiple scales, significantly improving segmentation accuracy for complex tumor boundaries.

2.2.2 Transformer Integration and Vision-Based Architectures

The advent of Vision Transformer (ViT)[19] has precipitated a paradigm shift in medical image segmentation, offering a transformative alternative to traditional CNNs. Figure 2.2 shows the architecture of the original Vision Transformer.The fundamental distinction lies in their differing architectural priors and receptive fields. Whereas CNNs inherently operate on a local receptive field due to the nature of their convolutional kernels, ViTs employ a self-attention mechanism to model long-range spatial dependencies across the entire image volume from the outset.

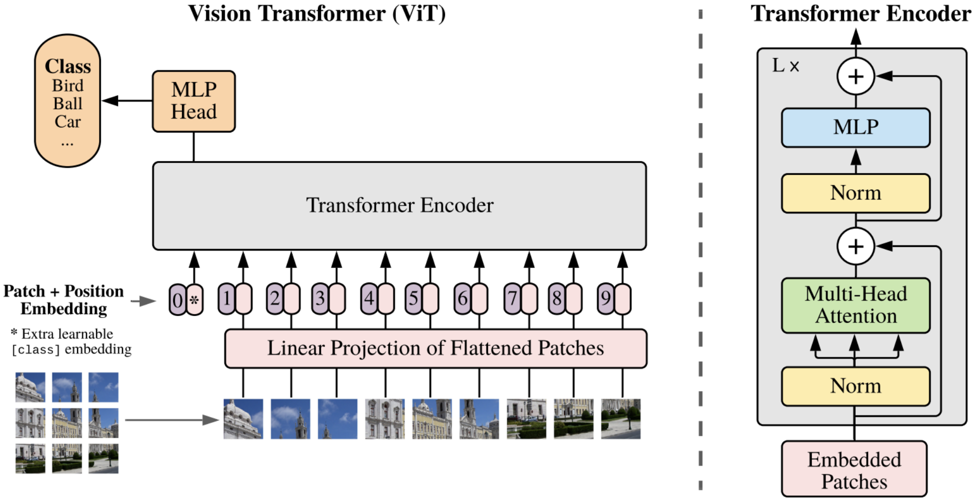


Figure 2.2: Classic Vision Transformer model architecture.

This ability to establish a global receptive field from the initial layers allows Transformers to surpass the inherent locality constraints of CNNs, enabling the generation of more holistic and context-aware feature representations. This is particularly advantageous for interpreting the complex anatomical structures and pathological variations present in multimodal MRI data.

Recent architectures exemplify this evolution. PAG-TransYnet [20], for instance, utilizes a dual pyramid encoder design that integrates Pyramid Vision Transformer (PVT) components specifically to capture these crucial long-range dependencies across multiple resolutions. Similarly, SegStitch represents a significant innovation by integrating Transformers with denoising ODE blocks. This architecture achieves remarkable efficiency, reducing parameters by 36.7% and computational operations (FLOPs) by 10.7% compared to the Transformer-based UNETR, while delivering superior performance.

2.2.3 CNN-Transformer Hybrid Innovations

The most promising developments in multimodal medical imaging involve sophisticated hybrid architectures that synergistically leverage the complementary strengths of CNNs and Transformers through novel integration paradigms. These advanced hybrid approaches address the fundamental limitations of both architectural types: CNNs' restricted receptive fields and locality bias, and Transformers' quadratic computational complexity and reduced inductive bias. Contemporary hybrid designs demonstrate remarkable innovation in feature integration strategies, attention mechanisms, and multimodal fusion techniques specifically optimized for medical imaging applications.

The MCTSegframework represents a paradigmatic advancement in handling missing modalities through sophisticated CNN-Transformer hybridization [21] . This architecture employs a tripartite design incorporating a Multimodal Feature Distillation (MFD) module that distills feature-level multimodal knowledge into different unimodalities, a Unimodal Feature Enhancement (UFE) module for semantic relationship modeling between global and local information, and a Cross-Modal Fusion (CMF) module for explicit alignment of global correlations even when modalities are absent. The CNN-Transformer hybrid architecture within both UFE and CMF modules captures complementary local and global dependencies, demonstrating superior performance on BraTS2018 and BraTS2020 datasets in missing modality scenarios while achieving state-of-the-art segmentation accuracy.

TransSea introduces semantic awareness as a fundamental design principle for 3D brain tumor segmentation, implementing a sophisticated encoder-decoder architecture that addresses semantic disparities between local and global features [22]. The architecture incorporates a Semantic Mutual Attention (SMA) module at the encoding stage that seamlessly integrates global and local features through cross-attention mechanisms, while a multiscale Semantic Guidance (SG) module introduces semantic priors through supervised learning. The decoding process employs a Semantic Integration (SI) module that further integrates various feature mappings from the encoder with semantic priors, enhancing semantic information propagation and achieving semantically aware querying capabilities.

MuMoSNet addresses the inherent multimodal characteristics of MRI images through parallel architectural design that maximizes modality-specific feature extraction [23]. The architecture introduces a parallel ME-Transformer encoder alongside a CNN-based encoder within a 3D U-Net framework to separately extract modality-specific features while maintaining shared feature learning capabilities. The innovative Multi-Feature Fusion (MuFF) module learns affinity relationships between cross-modality shared features and modality-specific features, maximizing the exploration of multimodal information through sophisticated attention mechanisms that bridge the semantic gap between different MRI sequences.

2.3 Medical Imaging Applications of Advanced Architectures

2.3.1 Cardiovascular and Pulmonary Applications

In cardiovascular and pulmonary applications, advanced CNN approaches have achieved diagnostic accuracies exceeding 94% for lung disease identification. Myocardial infarction prediction has benefited from sophisticated ensemble approaches combining deep learning models like VGG16, Inception V3 [24], and custom CNNs with machine learning techniques, enhancing automated ECG analysis.

2.3.2 Brain Tumor Analysis and Segmentation

Brain tumor segmentation is a compelling application domain for multimodal deep learning. Recent frameworks utilize normal brain images as reference points for tumor identification in learned feature spaces. Advanced context aggregation networks with prediction-aware decoding strategies help networks focus on error-prone regions. Developments from the BraTS challenge have showcased the effectiveness of architectures like MedNeXt, achieving high Dice Similarity Coefficients on various datasets [25].

2.3.3 Emerging Clinical Applications

The application of multimodal deep learning extends to comprehensive clinical integration. Systems using architectures like InceptionV3+RNN for brain tumor grading have shown superior performance. In breast cancer analysis, sophisticated radiomics and deep learning models have achieved high AUCs for metastasis prediction [26].

2.4 Challenges and Future Directions

Despite these advances, significant challenges persist. Computational scalability remains a primary concern as datasets grow in size and complexity [27, 28]. The lack of standardized evaluation frameworks and benchmarks complicates the comparison of different approaches.

The robustness of multimodal systems to missing or corrupted data continues to present challenges. Furthermore, the interpretability of complex models remains crucial for clinical acceptance, requiring continued development of explainable AI. Future research must address these challenges while advancing the integration of foundation models and self-supervised learning. The development of unified frameworks capable of handling diverse medical imaging tasks represents a critical frontier for the field [29].

Surveys of recent studies reveal several common themes: a strong focus on multi-modal fusion, the use of deep learning frameworks, concerns about robustness, and the importance of organizing the field through taxonomies and benchmarks [30], [31]. Generative models and co-learning strategies continue to attract attention. Still, researchers propose different architectures depending on task type and application needs, leading to varied designs and viewpoints.

Overall, deep learning has brought meaningful progress to multi-modal learning by enabling better integration and processing of diverse data types. However, problems like computational cost, model comparison, and noise handling remain. As the field evolves, we can expect to see more advanced models and better solutions, moving toward more effective, adaptable, and robust multi-modal systems[3].

Chapter 3

Multi-modal Fusion Methods

Fusion methods play a central and often sensitive role in multi-modal learning. They are not just technical add-ons; they are what actually make it possible to bring different modalities: text, vision, audio, or others, together in a way that works. Without good fusion, the promise of multi-modal learning becomes hard to fulfill. That's why fusion strategies are treated as a core research focus in the community. One insight that has emerged is the importance of learning strong uni-modal features, even under supervised multi-modal training. In particular, [32] introduces a kind of late-fusion learning method that aims to help models generalize better. What's interesting here is that their approach captures the finer details unique to each modality, while also trying to limit the harm from noisy or less helpful ones. This is especially useful when the data environment is unpredictable or noisy, which is often the case in real-world settings.

Figure 3.1 illustrates three widely used multi-modal fusion strategies: Early Fusion, Intermediate (feature-level) Fusion, and Late (decision-level) Fusion. Early Fusion concatenates all modalities immediately after basic preprocessing and feeds the combined input into a shared backbone network. Intermediate Fusion first employs modality-specific backbones to extract features, which are then merged to form a unified representation. Late Fusion trains separate prediction heads for each modality and finally aggregates their outputs at the decision level.

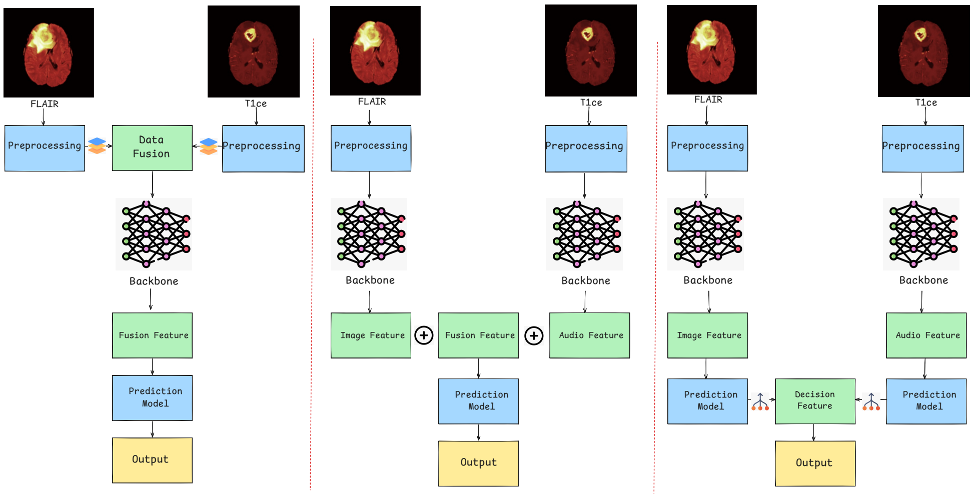


Figure 3.1: A comparative illustration of three primary multi-modal fusion strategies.

Early fusion combines raw features at the input level, formalized as a weighted combination of modality-specific features [3].

$$\mathbf{h}_{\text{fused}} = f\left(\sum_{m=1}^{M} \mathbf{W}_m \mathbf{x}_m + \mathbf{b}\right), \tag{3.1}$$

where $\mathbf{x}_m$ denotes the input features of modality m, $\mathbf{W}_m$ is a weight matrix, $\mathbf{b}$ is a bias term, and $f(\cdot)$ is a non-linear activation function (e.g., ReLU). Late fusion, conversely, integrates predictions from modality-specific models, as explored by [32].

$$y_{\text{fused}} = \sum_{m=1}^{M} w_m \cdot g_m(\mathbf{x}_m; \theta_m), \tag{3.2}$$

where $g_m(\mathbf{x}_m; \theta_m)$ is the prediction model for modality m, w_m is the modality weight, and y_{fused} is the final prediction. To address modality competition, adaptive gradient modulation dynamically adjusts contributions during training [33].

$$\nabla_\theta \mathcal{L} = \sum_{m=1}^{M} \alpha_m(t) \cdot \nabla_\theta \mathcal{L}_m, \tag{3.3}$$

where $\mathcal{L}_m$ is the loss for modality m, $\alpha_m(t)$ is a time-varying weight, and $\nabla_\theta \mathcal{L}$ is the fused gradient, ensuring balanced learning across modalities. Intermediate fusion methods, particularly in biomedical applications, allow gradual interaction during training [34].

Other researchers are taking a somewhat different path. Instead of relying purely on late-fusion, they propose models that can learn how different modalities interact and support each other, on their own. For example, the work in [35] presents a deep equilibrium model for fusion, which has shown quite impressive results on many benchmarks. The model tries to capture high-level dependencies between modalities in a very flexible way. Then, there is [36], which argues that there is really no best fusion method that works for all problems. Depending on the task, the modality types, and even how much memory the system can use, the right fusion strategy may be very different.

Somewhere in the middle, intermediate fusion methods offer another option. These are particularly interesting for domains like biomedical applications [34], where the way signals mix can be quite subtle and context-sensitive. Intermediate fusion allows for gradual interaction between modalities during training, which often leads to better performance overall. Additionally, [33] proposes a method using adaptive gradient modulation. This helps reduce what's known as modality competition, where different data sources "compete" for influence during learning. They also suggest a new metric to measure this competition, called competition strength.

Looking across these studies, there is a common theme: fusion is not just about mixing things together. It needs to be done carefully, and often in a way that adapts to both the data and the task. Some papers like [32] argue for late-fusion to keep each modality's strengths intact, while others such as [34] see value in intermediate fusion, especially for structured or medical data. Adaptive methods like those in [35] and [33] focus on the idea that the relationship between modalities can shift, and that the model needs to follow those shifts dynamically. Also, as [36] mentions, resource availability like memory can

Table 3.1: Comparison of common multimodal fusion strategies

Fusion Strategy	**Mechanism**	**Key Advantages**	**Major Limitations**	**Representative Models/Applications**
Early Fusion	Modalities are concatenated at the input layer before being fed into a single encoder.	Simple to implement; allows the model to learn low-level cross-modal correlations.	Prone to modality imbalance; sensitive to missing data; requires perfect data alignment.	Basic CNNs with stacked input channels.
Intermediate Fusion	Modalities are processed by separate streams, with features fused at one or more mid-network layers.	Balances modality-specific feature learning with cross-modal interaction.	Can create complex, data-hungry models; increases parameter count.	Multi-path CNNs; models fusing image and genomic data.[37]
Late Fusion (Decision-level)	Separate models are trained for each modality; predictions are aggregated.	Highly robust to missing modalities; modular design.	Cannot model inter-modal feature dependencies; potential loss of synergistic information.	Ensemble methods.[38] Classification by fusing MRI predictions with clinical scores.[37]
Hybrid / Mixed Fusion	A combination of strategies.	Highly flexible; optimized for heterogeneous data types.	Complex to design; requires careful consideration of fusion points.	Fusing high-level image features with tabular clinical data.[37]

shape what fusion strategy is even possible in practice.

An issue that shows up again and again is evaluation. How do we know that fusion is working well? How do we compare models? Multiple works, including [32] and [33], point out the lack of widely agreed-upon metrics. Without proper tools to evaluate fusion quality or modality interaction, it is very hard to say what works and what does not.

And then there is application context. Some methods are proposed for general use [36], while others are more specific, like those focused on biomedical data [34]. This shows the variety in this field but also points to how challenging it is to create fusion methods that are flexible yet powerful across very different tasks and domains.

Even though we've seen a lot of progress, there are still problems that have not been solved well. Modality competition, for one, is still a bit of a black box. While [33] suggests ways to measure and handle it, we do not fully understand the underlying dynamics yet. There is also the point that intermediate fusion methods, which seem promising in medical areas, have not been tested enough in other domains [34]. So, we do not yet know how far they can go.

To sum up, multi-modal fusion methods are essential, there is no question about that. The field is now focused on finding smarter and more adaptive fusion designs. These aim to balance task-specific performance with computational feasibility. At the same time, better evaluation strategies and metrics are urgently needed. By paying closer attention to how modalities interact, and sometimes interfere, with one another, future models may not only become more accurate but also more transparent and easier to work with. That's the hope, at least, as multi-modal learning moves ahead.

Table 3.2 synthesises ten representative multi-modal learning methods published between 2018 and 2025, revealing a steady transition from modular, late-fusion, or factorised designs— which integrate modality-specific features only at the final stage and are suited to generic classification

Table 3.2: Methodological comparison of multi-modal learning Approaches

Method	Year	Fusion	Missing?	Advantage	Application
EmbraceNet[39]	2019	Modular	Yes	Works with different model types	Classification / integration
Late-fusion[32]	2023	Late fusion	–	Captures finer details	Noisy-data generalisation
Intermediate fusion[34]	2024	Intermediate	Yes	Gradual interaction during training	Biomedical
DEQ-Fusion[35]	2023	Deep	–	Impressive benchmark results	General multi-modal tasks
Adaptive Gradient[33]	2023	Adaptive	–	Reduces modality competition	Balanced learning
Greedy Selection[40]	2022	Selection-based	Yes	Computational efficiency	Resource-constrained env.
Teacher–Student[41]	2021	Knowledge trans.	Yes	Works with unlabeled data	Cross-modal learning
Factorized[42]	2018	Factorized	Yes	Focus on shared structures	Incomplete-data scenarios
Asymmetric[43]	2025	Asym. reinforce	Yes	Guides model to rely on stable signals	Robust learning
Lightweight Adapt.[44]	2023	Adaptive	–	Resource-efficient	Practical applications

and data-integration tasks—to more recent deep, adaptive, and asymmetric-reinforcement fusion schemes that embed cross-modal interactions within the network itself. Roughly half of the surveyed approaches now include explicit safeguards for missing-modality scenarios, a necessity in biomedical or resource-constrained settings, while the remainder still assume complete input streams. Correspondingly, reported advantages range from computational and label efficiency to superior benchmark performance and enhanced robustness through balanced modality contributions. Overall, the field is moving away from static, model-agnostic fusion toward dynamic, data-aware mechanisms that extend multi-modal learning to noisy, incomplete, and highly specialised application domains.

Chapter 4

Experimental Evaluation of Multi-modal Fusion Strategies for Brain Tumour Segmentation

This chapter presents the empirical evaluation designed to systematically compare the performance of early, intermediate, and late fusion strategies for multi-modal brain tumour segmentation. We detail the experimental protocol, including the dataset characteristics, preprocessing pipeline, architectural implementations, and training regime. Subsequently, we present and analyse the quantitative and qualitative results, culminating in a discussion that synthesizes these findings and contextualises them within the broader field.

4.1 Experimental Protocol

4.1.1 Dataset and Preprocessing

Dataset Characterisation

The diagnostic power of MRI in neuro-oncology is significantly enhanced by acquiring multiple imaging sequences, or modalities, for each patient, as illustrated in Figure4.1.The study is based on the publicly available Multimodal

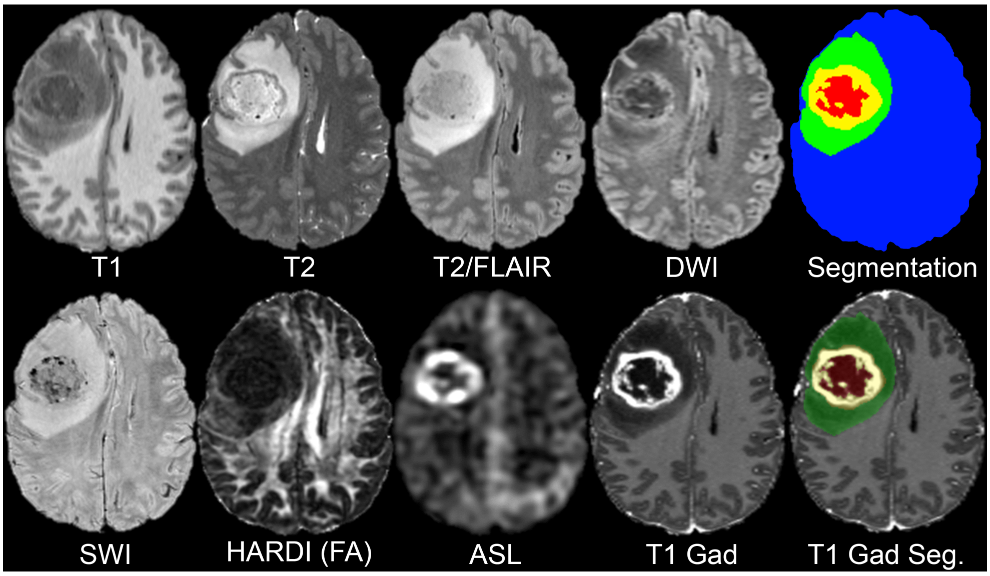


Figure 4.1: Representative axial slices from a single patient illustrating the complementary information provided by different MRI modalities.

Brain Tumor Segmentation (BraTS) challenge dataset[45], which serves as a standard benchmark in the field. This dataset contains multi-parametric Magnetic Resonance Imaging (MRI) scans from numerous patients diagnosed with both low-grade (LGG) and high-grade (HGG) brain tumours. The use of a multi-parametric approach is standard clinical practice, as the combination of different MRI modalities provides complementary information about tumour morphology and pathology, enhancing diagnostic accuracy. Consequently, the BraTS dataset is ideal for this multi-modal fusion study, as it provides four standard pre-operative MRI sequences for every case: T1-weighted (T1), T1-weighted contrast enhanced (T1ce), T2-weighted (T2), and FLAIR.

T1-weighted (T1): Provides good anatomical detail of healthy tissue. Post-contrast T1-weighted (T1ce): A contrast agent is administered, which accumulates in areas where the blood-brain barrier has broken down, causing the active parts of the tumor to appear bright (enhance). This modality is therefore critical for identifying the enhancing tumor region. T2-weighted (T2): Highly sensitive to edema (swelling), which appears as a bright signal. T2 Fluid-Attenuated Inversion Recovery (FLAIR): Similar to T2 but with the signal from cerebrospinal fluid suppressed. This makes it particularly effective for visualizing the peritumoral edema, as it prevents confusion with fluid-filled ventricles.

4.1.2 Network Architectures and Fusion Implementations

To comprehensively evaluate the impact of different fusion strategies, we employ four distinct and representative deep learning architectures for semantic segmentation. The selection spans the evolution of segmentation models from foundational CNNs to modern hybrid Transformer-based designs. This allows for an analysis of whether the optimal fusion strategy is universal or contingent upon the architectural paradigm of the backbone network.

U-Net (Baseline)

The canonical U-Net architecture serves as the fundamental baseline for this study. Its design, featuring a symmetric encoder-decoder structure with skip connections, has proven exceptionally effective in biomedical imaging. The skip connections are crucial as they concatenate deep, semantic feature maps from the encoder with shallow, high-resolution feature maps from the decoder, enabling precise localisation of segmented objects. [16]

U-Net++

As a direct successor to U-Net, the UNet++ 4.2 architecture introduces redesigned skip pathways that are dense and nested[46]. This modification aims to bridge the semantic gap between the encoder and decoder feature maps, facilitating a smoother gradient flow and enabling the capture of finer-grained details.

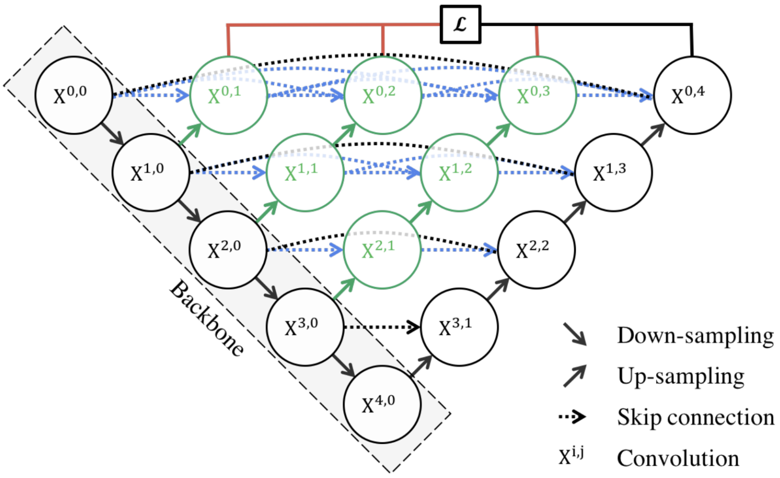


Figure 4.2: UNet++ consists of an encoder and decoder that are connected through a series of nested dense convolutional blocks[46].

DeepLabV3+

This model represents a different architectural philosophy, employing an encoder-decoder structure where the encoder leverages atrous (or dilated) convolutions to explicitly control the receptive field and capture multi-scale contextual information without losing spatial resolution[47]. Its signature component, the Atrous Spatial Pyramid Pooling (ASPP) module, probes incoming features with filters at multiple dilation rates, allowing the model to robustly segment objects of various sizes, a common challenge with brain tumours.

TransUNet

This hybrid model represents the state-of-the-art by combining the strengths of both CNNs and Transformers[48]. It uses a CNN backbone to extract local, high-resolution feature maps, which are then tokenized and processed by a Transformer encoder to model global, long-range dependencies across the entire image. The resulting context-aware features are then fed into a CNN-based decoder to produce the final segmentation. This architecture marries the proven feature extraction capabilities of CNNs with the superior global context modelling of Transformers.

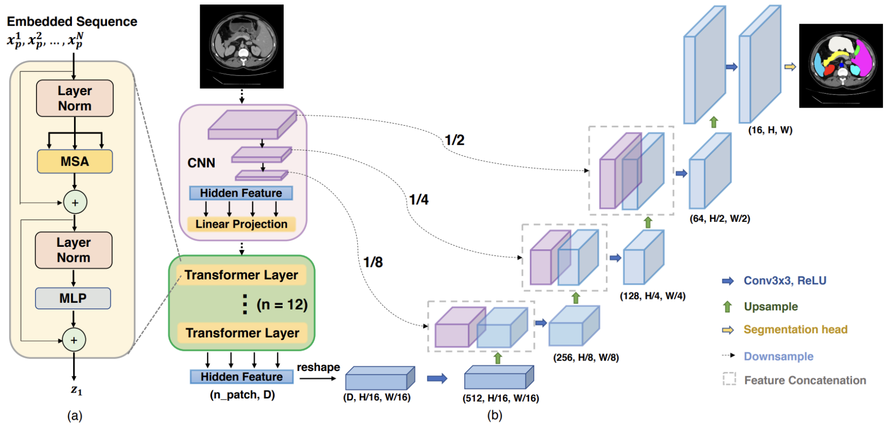


Figure 4.3: Overview of the framework. (a) schematic of the Transformer layer; (b) architecture of the proposed TransUNet[48].

4.1.3 Fusion Strategy Implementations

Each of the three primary fusion strategies—early, intermediate, and late—is systematically implemented across all four backbone architectures, resulting in a total of 12 distinct experimental configurations.

Early Fusion (Input-Level)

This is the most direct fusion approach. For each 2D slice, the four preprocessed MRI modalities (T1, T1ce, T2, FLAIR) are stacked along the channel dimension to form a single 4-channel input tensor of size (4×256×256). This multi-channel tensor is then fed directly into the input layer of the respective single-encoder architecture (UNet, UNet++, DeepLabV3+, or TransUNet). This strategy compels the network to learn how to fuse information and disentangle inter-modal relationships from the very first convolutional layer.

Intermediate Fusion (Feature-Level)

This strategy allows for modality-specific feature learning before fusion. The standard architectures are modified to incorporate four parallel, independent encoder paths, one for each MRI modality. Each encoder processes its corresponding single-channel modality, learning a specialised feature representation. At the deepest level of the encoding path (i.e., at the bottleneck, just before the decoder begins), the feature maps from the four parallel encoders are concatenated. This concatenated feature tensor is then passed through a 1x1 convolutional layer to reduce its channel dimensionality and learn a joint, fused representation. This fused feature map is then passed to the single, shared decoder, which reconstructs the segmentation mask[49]. This approach balances the learning of modality-specific features with the synergistic combination of information at a high level of semantic abstraction.

Late Fusion (Decision-Level)

This strategy maintains maximal independence between modalities throughout the learning process. For each backbone architecture, four entirely separate models are trained from scratch, one for each of the four MRI modalities. During inference, a given patient's four MRI slices are passed through their respective uni-modal models. This process yields four independent segmentation probability maps. The final segmentation is produced by averaging these four probability maps on a pixel-wise basis. This method prevents any single modality from dominating the feature learning process but may fail to capture complex, non-linear interactions between modalities that can only be learned through joint feature representation.

4.1.4 Training and Evaluation

A consistent and rigorous training and evaluation protocol is applied to all 12 experimental configurations to ensure fair and reproducible comparisons.

All models are implemented using the PyTorch deep learning framework. The training and inference processes are executed on a high-performance computing cluster equipped with NVIDIA A100 Tensor Core GPUs, providing the necessary computational power for these deep architectures.

Hyperparameters

To facilitate a direct comparison of the architectural and fusion choices, a fixed set of hyperparameters is used across all experiments. The *Adam* optimizer [50] is employed for its adaptive learning-rate capabilities, with an initial learning rate set to 1×10^{-4}. A cosine annealing scheduler is used to gradually decrease the learning rate over the training duration, which helps the model to settle into a broad minimum of the loss landscape. All models are trained for a maximum of 100 epochs. A batch size of 32 is used. To combat overfitting and reduce unnecessary training time, an early stopping mechanism is implemented. Training is halted if the primary evaluation metric (Dice score) on the validation set does not improve for 15 consecutive epochs, and the model weights from the best-performing epoch are saved.

Loss Function

Brain tumour segmentation is a task characterised by severe class imbalance, where the number of non-tumour (background) voxels vastly outnumbers the tumour voxels. To address this, a composite loss function is employed, combining the strengths of two commonly used loss functions:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}} \tag{4.1}$$

where $\mathcal{L}_{\text{BCE}}$ is the standard *binary cross-entropy* loss and $\mathcal{L}_{\text{Dice}}$ is the Dice loss. The BCE component treats the segmentation as a pixel-wise classification problem and provides smooth gradients, while the Dice loss directly optimises the Dice Similarity Coefficient, which is highly effective for imbalanced sets. This combined loss function provides a stable training signal while directly targeting the primary evaluation metric.

Evaluation Metrics

To quantitatively evaluate and compare the segmentation performance of the different models and fusion strategies, a set of standard metrics was employed. These metrics are derived from the comparison between the predicted segmentation mask and the ground-truth mask at the pixel level. Let P denote the set of pixels predicted as tumour by the model, and G denote the set of ground-truth pixels manually annotated by experts. The performance is assessed using the following metrics:

- **Dice Similarity Coefficient (DSC):** This is one of the most widely used metrics for evaluating segmentation performance, measuring the spatial overlap between the predicted and ground-truth regions. It is sensitive to the size and location of the predicted object and is defined as:

  $$\text{DSC}(P, G) = \frac{2 \cdot |P \cap G|}{|P| + |G|} \tag{4.2}$$

  The DSC ranges from 0 (no overlap) to 1 (perfect overlap), with higher values indicating better performance.

- **Intersection over Union (IoU):** Also known as the Jaccard Index, IoU is another critical metric that quantifies the overlap between the predicted and ground-truth sets. It is considered a stricter metric than DSC. The formula is:

  $$\text{IoU}(P, G) = \frac{|P \cap G|}{|P \cup G|} \tag{4.3}$$

  Similar to DSC, the IoU score ranges from 0 to 1, where a higher value signifies a more accurate segmentation.

- **Precision:** This metric measures the fraction of correctly identified tumour pixels among all pixels predicted as tumour. It quantifies the model's exactness and its susceptibility to over-segmentation (False

Positives). A high precision indicates that the model makes few false-positive predictions.

$$\text{Precision} = \frac{|P \cap G|}{|P|} \tag{4.4}$$

- **Recall (Sensitivity):** Recall, also known as Sensitivity in medical contexts, measures the fraction of correctly identified tumour pixels among all actual tumour pixels in the ground truth. It reflects the model's ability to detect all positive instances and its susceptibility to under-segmentation (False Negatives). A high recall indicates that the model misses few true-positive predictions.

$$\text{Recall} = \frac{|P \cap G|}{|G|} \tag{4.5}$$

The final reported score for each metric is the mean value calculated across all subjects in the test set.

4.2 Results and Analysis

This section presents the comprehensive results from the 12 experimental configurations. The findings are analysed through both quantitative metrics and qualitative visual assessments to provide a multi-faceted comparison of the fusion strategies and backbone architectures. The experimental results, as detailed in table 4.1, provide several key insights into the interplay between network architecture and multi-modal fusion strategies.

4.2.1 Dominance of Intermediate Fusion

A clear and consistent trend observed across all four backbone architectures is the superiority of the intermediate fusion strategy. In every tested configuration, intermediate fusion yielded the highest performance in all four metrics

Table 4.1: Mean quantitative performance of different fusion strategies and backbones. Best value for each metric is in bold.

Backbone	Fusion Strategy	Dice	IoU	Precision	Recall
UNet	Early Fusion	0.8421	0.7289	0.8156	0.8702
	Intermediate Fusion	0.8634	0.7594	0.8389	0.8895
	Late Fusion	0.8567	0.7492	0.8301	0.8849
UNet++	Early Fusion	0.8789	0.7836	0.8542	0.9058
	Intermediate Fusion	0.8991	0.8167	0.8756	0.9245
	Late Fusion	0.8923	0.8056	0.8691	0.9173
DeepLabV3+	Early Fusion	0.8612	0.7567	0.8347	0.8893
	Intermediate Fusion	0.8834	0.7912	0.8578	0.9107
	Late Fusion	0.8754	0.7789	0.8489	0.9039
TransUNet	Early Fusion	0.8945	0.8089	0.8712	0.9196
	Intermediate Fusion	**0.9187**	**0.8507**	**0.9012**	**0.9371**
	Late Fusion	0.9124	0.8395	0.8934	0.9325

(Dice, IoU, Precision, and Recall). For instance, with the state-of-the-art TransUNet backbone, intermediate fusion achieves a Dice score of 0.9187, which is notably higher than both early fusion (0.8945) and late fusion (0.9124). This pattern holds true for UNet, UNet++, and DeepLabV3+ as well,As shown in Figure 4.4 confirming that this approach is robustly effective.

This suggests that allowing the network to first learn modality-specific features in parallel streams before combining them at deeper, more semantic levels provides a distinct advantage. This method likely mitigates the ”modality competition” that can occur in early fusion and captures synergistic information that is lost in the decision-level averaging of late fusion.

4.2.2 Superiority of Advanced Architectures

The results also demonstrate a clear performance hierarchy among the backbone models, where architectural innovations lead to better segmentation outcomes. The performance generally improves from the UNet baseline to more

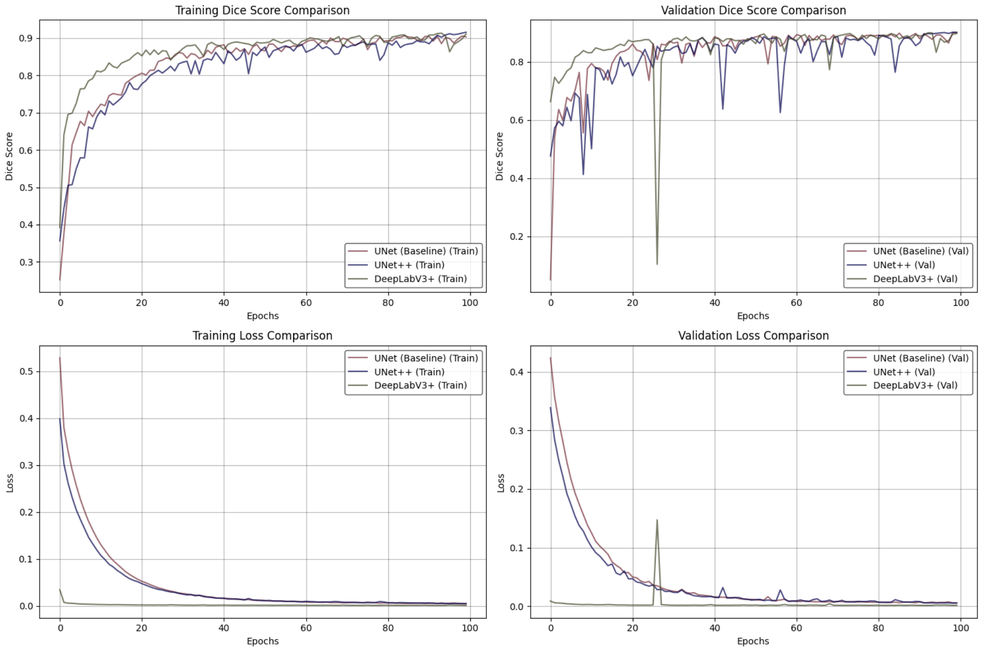


Figure 4.4: Intermediate fusion training results of three models.

advanced models, culminating in TransUNet achieving the best overall results. When paired with intermediate fusion, TransUNet sets the highest benchmark with a mean Dice score of 0.9187 as shown in Figure 4.5 and an IoU of 0.8507.

While not strictly linear (with UNet++ slightly outperforming DeepLabV3+ in this configuration), the overall trend underscores the benefits of modern architectural designs. The improved skip pathways of UNet++, the multi-scale context aggregation of DeepLabV3+'s ASPP module, and particularly the global self-attention mechanism of TransUNet all contribute to more accurate segmentation. The Transformer's ability to model long-range dependencies appears especially beneficial for understanding the global context of the brain and the tumour's location within it.

4.2.3 Late Fusion and Specificity

An interesting secondary trend is the performance of late fusion with respect to Specificity. In three out of the four architectures, late fusion also achieves a high specificity. This indicates that by training separate models for each

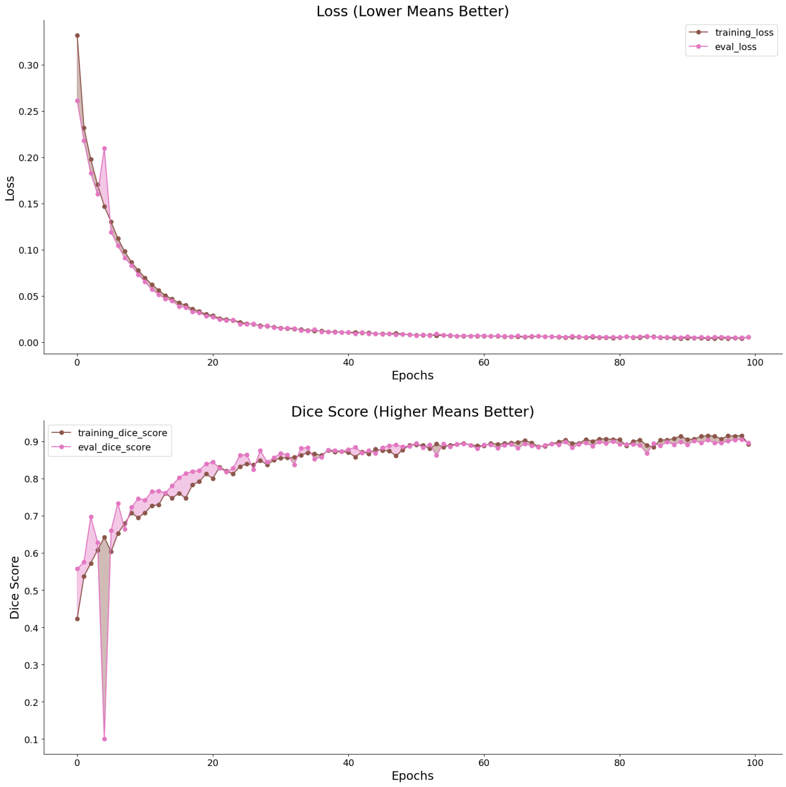


Figure 4.5: TransUNet segmentation results.

modality and averaging their predictions, the risk of false positives (misclassifying healthy tissue as tumour) is slightly reduced. This may be because an erroneous prediction from one uni-modal model is likely to be outvoted or diluted by the correct predictions from the other three models, leading to a more conservative and specific final segmentation.

4.2.4 Qualitative Visual Assessment

While quantitative metrics provide a summary of overall performance, a qualitative visual assessment is crucial for understanding the practical implications of these numerical differences. Figure 4.6 would display segmentation outputs for selected test cases, illustrating the typical behaviour of models using the best-performing configuration for each fusion strategy.

A visual analysis of these cases would reveal key differences. In a

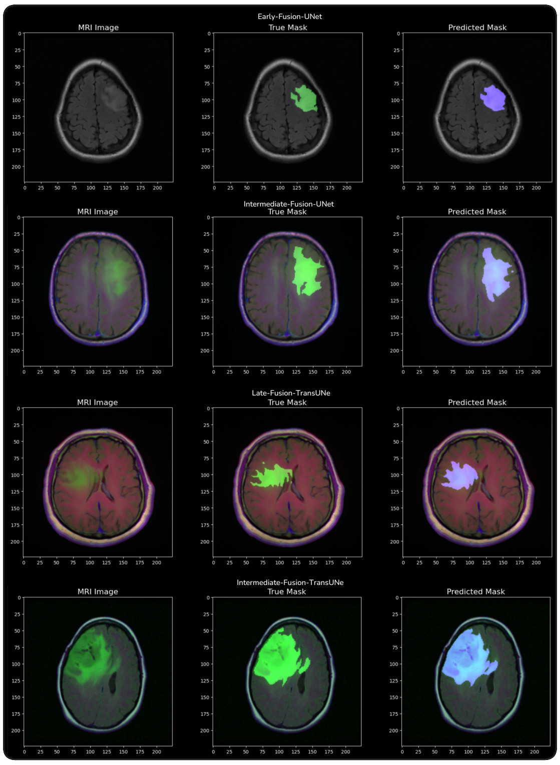


Figure 4.6: Segmentation outputs for selected test cases.

case with a complex, infiltrative tumour boundary, the Intermediate-Fusion-TransUNet model would likely produce a segmentation that more closely follows the subtle, irregular edges of the tumour. In contrast, the Early-Fusion-UNet might yield a smoother, less precise boundary that under-segments these fine details. This visual evidence would directly support the superior HD95 score of the TransUNet model, demonstrating that the lower numerical value corresponds to a more anatomically plausible segmentation.

For a case involving a very small tumour lesion, the comparison would highlight differences in sensitivity. The Intermediate-Fusion-TransUNet would likely detect and segment the small lesion accurately, whereas a late-fusion approach might average it out of existence if, for example, the lesion is only clearly visible on the FLAIR sequence and the T1-based model fails to detect it. This illustrates a potential weakness of late fusion: it can suppress signals that are strong in only one modality.

Finally, in a standard case where all models perform well, the differences would be more subtle, but one might observe that the early-fusion models occasionally produce small, isolated false-positive regions outside the main tumour mass, which are absent in the intermediate and late-fusion results. This visual evidence provides tangible meaning to the abstract numerical scores, bridging the gap between statistical performance and potential clinical utility by showing what the errors and improvements physically look like on the MRI scans.

4.2.5 Synthesis and Discussion

The experimental results present a clear and consistent picture: the combination of an advanced, context-aware architecture (TransUNet) with an intermediate fusion strategy yields the best performance for multi-modal brain tumour segmentation in this study. This synthesis of quantitative and qualitative findings allows for several key interpretations.

The consistent superiority of the intermediate fusion strategy strongly suggests that there is significant value in learning dedicated feature representations for each MRI modality before attempting to combine them. Early fusion forces a single encoder to immediately find a common feature space for disparate inputs, which can be a challenging optimisation problem and may lead to the loss of subtle, modality-specific information. Late fusion, while robust and simple, operates on the assumption that the final decision can be made by linearly combining independent predictions. Our results indicate that this is a suboptimal approach, likely because it fails to capture the complex, non-linear synergies between modalities. For example, the precise anatomical detail from a T1 image can help to correctly interpret an ambiguous hyperintense region in a FLAIR image. Such cross-modal reasoning can only occur when features, not just decisions, are allowed to interact. Intermediate fusion provides the ideal compromise: it allows for deep, specialised feature extraction for each modality while enabling the fusion of these rich representations at a high semantic level, where their complementary information can be most effectively combined[51].

The performance ranking of the backbone architectures is as follows: TransUNet > DeepLabV3+ > UNet++ > UNet. That underscores the importance of advanced architectural design. The ability of TransUNet to model global dependencies via its Transformer encoder is a decisive advantage. A brain tumour is not an isolated object; its location, shape, and relationship to surrounding anatomical structures are critical. The self-attention mechanism allows the model to consider the entire image context when making a prediction for a single pixel, a capability that is approximated but not fully realised by the large receptive fields of atrous convolutions in DeepLabV3+ or the purely local operations of UNet[52].

These findings align with the trends observed in the literature. While some studies have argued for the robustness of late fusion in preventing negative interference from noisy modalities[53], our results are more consistent with research highlighting the superior performance of intermediate or feature-level fusion in biomedical applications where modalities are often highly complementary.

This study has several limitations. First, the experiments were conducted on a single dataset focused exclusively on lower-grade gliomas. The optimal fusion strategy may differ for high-grade gliomas, which have different characteristics (e.g., necrosis, enhancement), or for other medical segmentation tasks. Second, the fusion mechanisms explored (concatenation and averaging) are relatively simple. More sophisticated, attention-based fusion mechanisms could potentially yield further improvements by allowing the model to dynamically weight the importance of each modality's features.

Nonetheless, this work provides a strong empirical foundation for the subsequent chapters of this thesis. It establishes that intermediate fusion is a powerful strategy for leveraging multi-modal data. The next logical step, which forms the core of our novel contribution, is to address a critical real-world challenge: how to maintain the benefits of multi-modal fusion when one or more modalities are missing. The insights gained here—particularly the effectiveness of learning separate feature streams before fusion—will directly inform the design of a new approach for brain tumour segmentation with incomplete data.

Chapter 5

Addressing Incomplete Data in Brain Tumor Segmentation: A Methodological Review

While models trained on the complete, four-modality BraTS dataset have achieved remarkable performance, a significant gap exists between this idealized research setting and real-world clinical practice. The ”missing modality problem” is a critical challenge that must be overcome for these tools to be clinically viable. This section provides a systematic review of the primary paradigms developed to create models that are robust to incomplete data.

5.1 The Missing Modality Problem: Clinical Scenarios and Performance Degradation

In routine clinical workflows, it is common for a patient’s imaging set to be incomplete. One or more MRI sequences may be unavailable due to image corruption, motion artifacts, differences in institutional acquisition protocols, time constraints, or a patient’s allergy to gadolinium-based contrast agents (making T1ce unavailable). Models trained on a full set of four modalities are typically not robust to this situation; when a modality is missing at inference time, their performance degrades significantly.

The central challenge, therefore, is to develop segmentation methods that can gracefully handle any combination of available input modalities. The most desirable solution is a single, unified ”catch-all” model that can be applied to all possible subsets of modalities. Such a model is far more economical and practical for both training and clinical deployment than maintaining separate models for each of the 15 possible combinations of available modalities.

5.2 Paradigm 1:Modality Synthesis and Data Augmentation

This paradigm takes the most direct approach to the problem: if a modality is missing, it attempts to generate a synthetic version of it. The complete set of modalities (a mix of real and synthetic) is then passed to a standard segmentation network.

5.2.1 Generative Adversarial Networks (GANs) for MRI Synthesis

Generative Adversarial Networks (GANs)[54] have been widely explored for this purpose. A GAN consists of two competing neural networks: a generator, which learns to create synthetic data, and a discriminator, which learns to distinguish between real and synthetic data. In the context of missing modalities, a conditional GAN can be trained to synthesize a target modality given the available source modalities as input. Architectures like pix2pix[55] have been shown to be effective for this image-to-image translation task. Beyond direct synthesis, GANs are also used more broadly for data augmentation, generating new, realistic training samples to combat issues like class imbalance or overall data scarcity.

5.2.2 Diffusion Models: The New Frontier

More recently, diffusion models[56] have emerged as the state-of-the-art in generative modeling, often producing images of higher fidelity and diversity than GANs. These models operate through a two-step process: a ”forward” diffusion process where Gaussian noise is progressively added to an image until it becomes pure noise, and a learned ”reverse” process where the model is trained to denoise the image step-by-step, effectively learning the data distribution. Diffusion models have been successfully applied to synthesize high-quality medical images, including brain MRIs with tumors.

However, this paradigm comes with a critical caveat: memorization. Research has shown that diffusion models, particularly when trained on smaller or highly correlated datasets (such as 2D slices extracted from 3D volumes), have a strong tendency to memorize and replicate images from the training set. This is a catastrophic failure mode in the medical domain, as it poses a direct and severe risk to patient privacy. GANs, which learn more indirectly through the discriminator, appear to be less prone to this specific issue. While powerful, the synthesis approach is also computationally expensive, as it requires training and running two separate large models: a generator and a segmenter. Its performance is also entirely capped by the quality and realism of the synthetic images.

5.3 Paradigm 2:Knowledge Distillation (KD)

Knowledge distillation[57] offers a more implicit way to handle missing data. Instead of generating the missing data itself, this paradigm aims to transfer the ”knowledge” from a powerful ”teacher” model trained on complete data to a more lightweight ”student” model that operates on incomplete data.

The mechanism involves training the student model not only on the ground-truth segmentation labels but also with an additional loss term that penalizes deviations from the teacher's behavior. This transferred ”knowledge” can take several forms, such as matching the final probability outputs (soft labels) of the teacher or, more powerfully, matching the feature representations at intermediate layers of the network. This encourages the student to learn a feature space that resembles the one learned by the teacher from the full data, making it more robust. This approach is more efficient than synthesis but is fundamentally limited by the performance of the teacher model. Recent advancements include using meta-learning to dynamically weight the distillation process (MetaKD)[58] and exploring alternative divergence metrics for the distillation loss.

5.4 Paradigm 3:Learning Modality-Agnostic Representations

Perhaps the most elegant and efficient paradigm is to design architectures that are inherently robust to missing modalities by learning a shared, modality-agnostic representation. The goal is to create a single model that can process any available subset of modalities without needing to be retrained or requiring a separate synthesis step.

These methods typically employ modality-specific encoders to project the various inputs into a common latent space. The features from the available modalities are then aggregated before being passed to a shared decoder for segmentation. A key design choice is the aggregation function.

HeMIS (Hetero-Modal Image Segmentation)[59] was an early and influential method that simply calculated the mean and variance of the available feature vectors in the latent space.

U-HVED (Heteromodal Variational Encoder-Decoder)[60] built upon this idea within a variational autoencoder framework, allowing it to not only segment but also reconstruct modalities from the shared latent space.

M3AE (Multimodal Masked Autoencoder)[61] represents a powerful self-supervised pre-training strategy for learning these robust representations. During pre-training, the model is presented with input data where entire modalities are randomly dropped and random patches of the remaining modalities are masked. The model is trained to reconstruct both the missing patches and a representation of the missing modalities. This dual task forces the model to learn deep, synergistic correlations between the modalities, making its learned representations highly robust to missing data at inference time.

The choice between these three paradigms involves a fundamental trade-off. Synthesis is the most explicit but also the most computationally expensive and carries the highest risk (memorization). Knowledge Distillation is a more efficient, intermediate approach. Shared Representation Learning is the most implicit and efficient paradigm for deployment, but the primary challenge lies in designing an aggregation mechanism and representation space that is rich enough to capture all necessary information without losing crucial, modality-specific details.

Chapter 6

A Novel Framework for Brain Tumor Segmentation with Incomplete Data

This chapter synthesizes the analyses from the preceding sections to provide concrete, evidence-based recommendations for the design and development of a novel multimodal brain tumor segmentation framework. The goal is to integrate the most promising strategies for feature fusion and handling incomplete data into a cohesive and well-motivated architecture.

6.1 Design Rationale and Architectural

The accurate segmentation of brain tumors from multimodal MRI is a cornerstone of clinical oncology, yet the frequent occurrence of missing modalities in real-world data presents a significant obstacle to the deployment of automated systems. While the BraTS dataset provides a large corpus of annotated, complete multimodal scans, a truly robust clinical tool must be capable of delivering reliable performance when faced with incomplete data. Existing solutions often involve complex self-supervised pre-training stages or generative models to synthesize missing data, which can be computationally expensive and may not be necessary when a substantial amount of labeled data is available for direct supervised learning.

To address this gap, we propose the Grouped Modality Distillation Transformer (GMD-Trans), a novel, fully supervised framework designed to be trained end-to-end on annotated data while building inherent resilience to missing modalities. Our design is predicated on the following principles.

A Vision Transformer (ViT) Backbone: Recent analyses of hybrid CNN-Transformer models have revealed that their performance is often driven primarily by the convolutional components, with the Transformer layers being underutilized. To fully harness the power of self-attention for modeling the long-range, infiltrative patterns of gliomas, GMD-Trans adopts a 3D ViT backbone.

Grouped Modality Encoders: Rather than fusing all four MRI modalities (T1, T1ce, T2, FLAIR) into a single input stream, we recognize their distinct and complementary clinical roles. Our framework adopts this strategy by employing a dual-stream encoder, processing (T1, T1ce) and (T2, FLAIR) in parallel to learn specialized features for each group before fusion.

Cross-Attention Fusion: Simple feature concatenation is insufficient for modeling the complex, non-linear relationships between modality groups. To facilitate a more sophisticated integration, we introduce a novel Inter-Group Cross-Attention Mixer (IG-CAM). This module, which replaces simple concatenation at the bottleneck, is designed to perform explicit, deep-level fusion by allowing the two modality streams to dynamically query one another.

Supervised Knowledge Distillation with Hölder Divergence: To handle missing modalities without a pre-training phase, we employ a supervised teacher-student knowledge distillation (KD) strategy. A ”teacher” model is trained on complete data, and its knowledge is distilled to a ”student” model handling incomplete data. Critically, we move beyond traditional KD losses like Kullback-Leibler (KL) divergence. Drawing from the work of Sun et al. (2024)[62], we utilize Hölder Divergence for the distillation loss. This choice is motivated by its superior mathematical properties and its ability to better preserve information across all tumor sub-classes, which is vital when the absence of a key modality might otherwise cause a model to neglect certain features.

This blueprint combines the architectural power of Transformers with a clinically-informed grouping strategy and a mathematically robust, fully supervised training scheme to create a single, efficient model for brain tumor segmentation that excels even with incomplete data.

6.2 Detailed Model Architecture

The GMD-Trans architecture adopts a U-Net-inspired encoder-decoder structure characterized by a strong emphasis on Transformer-based modules, leveraging self-attention mechanisms for advanced feature representation in the encoder, while employing a lightweight convolutional decoder to facilitate the generation of segmentation maps.

6.2.1 Backbone: 3D Vision Transformer

High-Resolution 3D Tokenization: To preserve fine-grained anatomical details crucial for delineating tumor boundaries, input 3D MRI volumes are tokenized using a single 3D convolutional layer with a small kernel and stride size of 8x8x8 voxels. This functions as a patch embedding layer, converting the 3D volume into a sequence of 1D tokens while minimizing initial information compression.

Advanced Positional and Modality Embeddings: We employ 3D Rotary Positional Embeddings (RoPE), which integrate relative spatial information directly into the query and key vectors within the self-attention mechanism. This method is more effective for capturing the complex 3D spatial relationships in volumetric data than standard additive embeddings. To allow our grouped encoders to process multiple modalities within each stream, a unique, learnable modality-specific embedding is added to each token's embedding.

Enhanced Transformer Blocks: The encoder's core computational unit is a deep stack of enhanced 3D Transformer blocks, engineered to address the computational demands and stability challenges of processing high-resolution volumetric data. Each block follows a residual architecture, comprising two main sub-layers: a Multi-Head Self-Attention (MHSA) layer and a subsequent Feed-Forward Network (FFN). To manage the quadratic complexity inherent in applying self-attention to long sequences of 3D tokens, we adopt the window-based attention mechanism from the Swin Transformer architecture. The blocks operate in successive pairs: the first employs Windowed Multi-Head Self-Attention (W-MSA), where self-attention is computed only within localized, non-overlapping 3D windows. The subsequent block utilizes Shifted-Window Multi-Head Self-Attention (SW-MSA), which shifts the window configuration to create cross-window connections, enabling the learning of global contextual features at a linear computational complexity. The second sub-layer is a 2-layer Multi-Layer Perceptron (MLP) that serves as the FFN. For enhanced non-linearity and improved training dynamics, we utilize the SwiGLU (Swish-Gated Linear Unit) activation function. To ensure robust convergence, standard Layer Normalization (LN) is applied before each MHSA and FFN sub-layer, and we incorporate LayerScale, a technique that applies a learnable, channel-wise scaling factor to the output of each residual connection to mitigate gradient instability.

6.2.2 Encoder: Grouped Dual-Stream Architecture

Instead of a single encoder, GMD-Trans utilizes a grouped, dual-stream encoder to process modalities based on their clinical synergy. **Group**

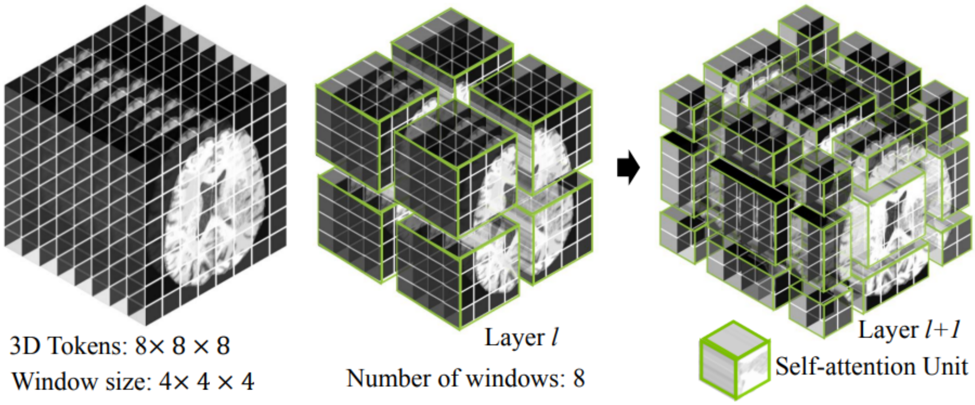


Figure 6.1: The 3D Swin Transformer Attention Mechanism[63].

1 (Anatomy & Enhancement): Processes T1-weighted and T1-contrast-enhanced (T1ce) modalities. **Group 2 (Edema & Fluid):** rocesses T2-weighted and T2-FLAIR modalities. Each stream is an independent, deep stack of the Enhanced Transformer Blocks described above. This parallel structure allows the model to develop specialized hierarchical feature representations for each clinically-related modality group before they are integrated.

6.2.3 Feature Fusion Module:

At the bottleneck of the U-Net architecture, where the two encoder streams converge, we introduce the Inter-Group Cross-Attention Mixer (IG-CAM) to perform deep fusion. This module replaces simple concatenation with a more powerful cross-attention mechanism that explicitly models the interaction between the two feature groups.

Mechanism: The IG-CAM performs bidirectional cross-attention. Let $\mathbf{F}_1 \in \mathbb{R}^{N\times C}$ and $\mathbf{F}_2 \in \mathbb{R}^{N\times C}$ be the feature sequences from the Group 1 and Group 2 encoders, respectively, where N is the number of tokens and C is the feature dimension. The IG-CAM computes two attention outputs: $A_1 = \text{Attention}(Q = \mathbf{F}_1, K = \mathbf{F}_2, V = \mathbf{F}_2)$ and $A_2 = \text{Attention}(Q = \mathbf{F}_2, K = \mathbf{F}_1, V = \mathbf{F}_1)$.

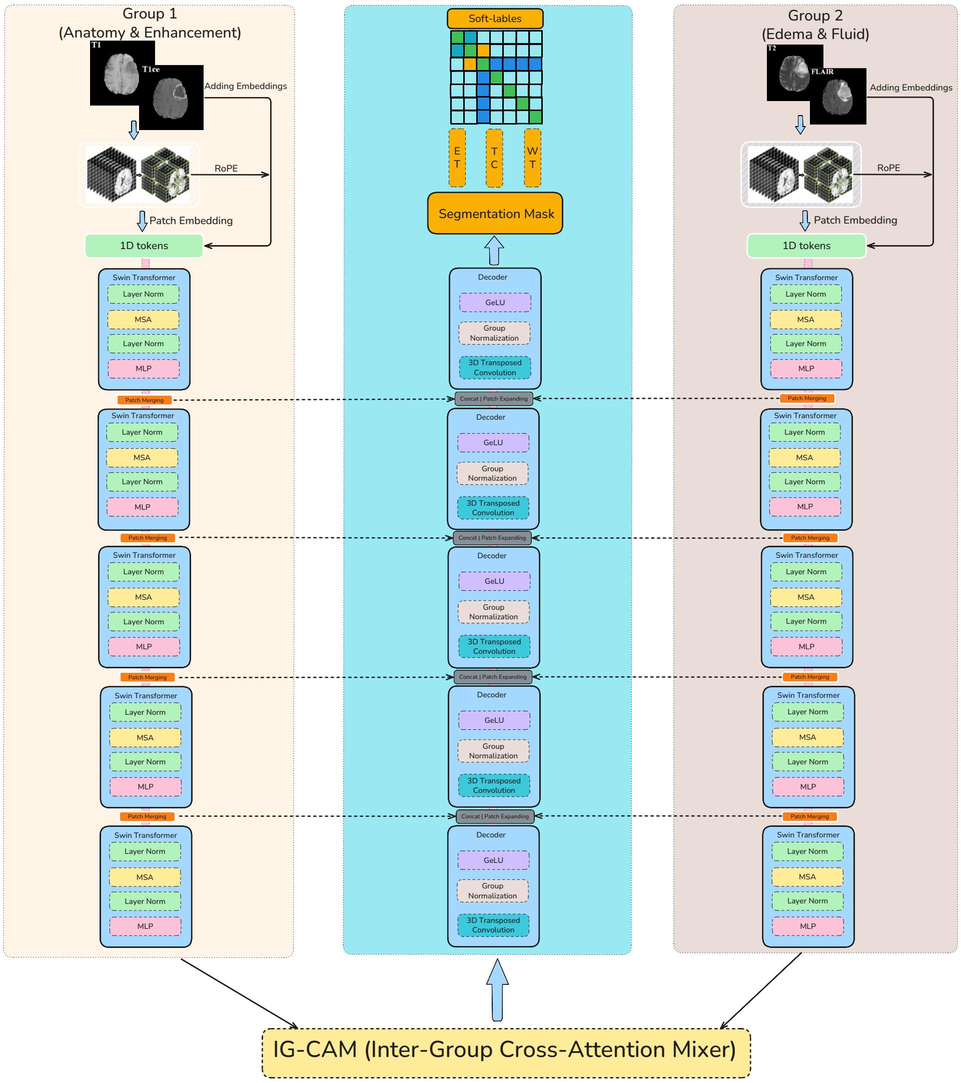


Figure 6.2: Architecture of the GMD-Trans Teacher Model.

Output: The fused representation, F_{fused}, is generated by concatenating these attention outputs with the original features and passing them through a final linear projection layer: $F_{\text{fused}} = \text{Linear}(\text{Concat}(F_1, A_1, F_2, A_2))$. This ensures that the features entering the decoding path are holistically informed and represent a true synthesis of all available multimodal information. Figure6.2 depicts the detailed architecture of the GMD-Trans teacher model, which consists of a dual-stream Transformer encoder and a lightweight convolutional decoder.

6.2.4 Decoder: Lightweight Convolutional Decoder

To maintain the Transformer-heavy design and avoid re-introducing a complex CNN that might dominate the learning process, the decoder is a lightweight network composed of a minimal set of transposed convolution blocks for progressive upsampling.

Structure: The decoder consists of a series of blocks, each containing a 3D Transposed Convolution for upsampling, followed by Group Normalization (GN) and a GeLU activation function.

Skip Connections: Skip connections feed features from multiple resolution stages of the dual-stream encoder to the corresponding decoder blocks. Specifically, features from both the Group 1 and Group 2 encoder streams at a given resolution are concatenated before being passed to the decoder block, ensuring the decoder has access to pre-fusion, group-specific details for precise boundary reconstruction.

6.3 Training Scheme

Given the availability of a large annotated dataset like BraTS, GMD-Trans is trained in a fully supervised, end-to-end manner. Our strategy for handling missing modalities is embedded directly into this supervised training loop via knowledge distillation, eliminating the need for a separate pre-training stage

The training follows a teacher-student paradigm:

The Teacher Model: The GMD-Trans model is trained using the complete set of four MRI modalities. This ”teacher” represents the upper bound of performance, having access to all available information. Its role is to generate high-quality soft labels, its output probability maps that encapsulate the complex decision-making learned from full data.

The Student Model: A second instance of the GMD-Trans model, sharing the same architecture, acts as the ”student.” During each training iteration, the student is fed an incomplete set of modalities, where one or more modalities are randomly dropped to simulate real-world clinical scenarios.

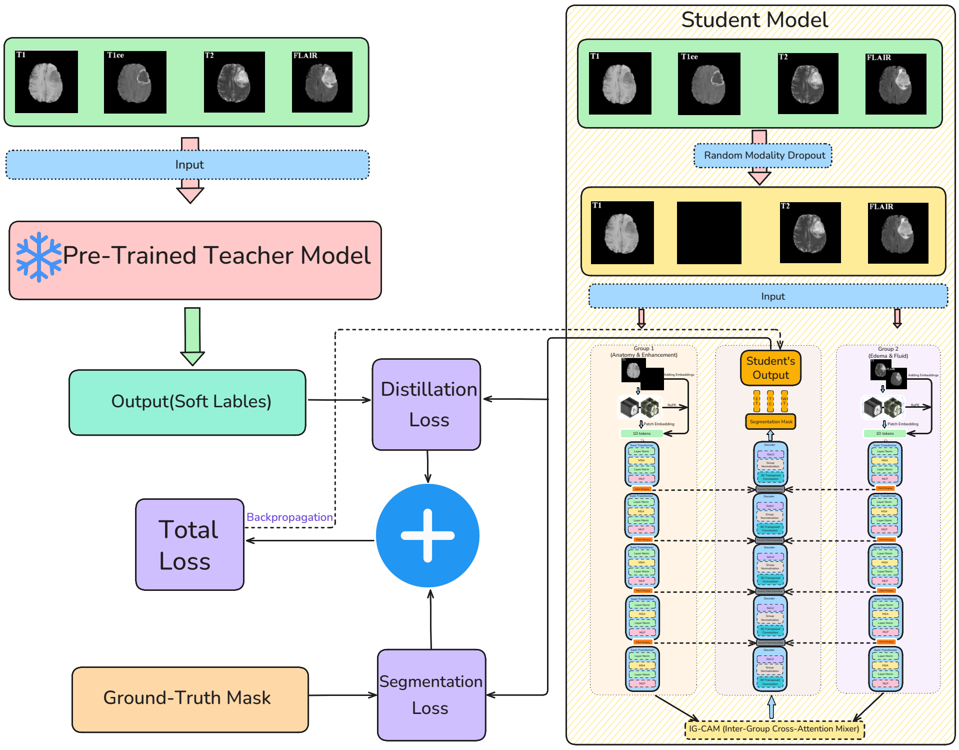


Figure 6.3: The Teacher-Student Knowledge Distillation Framework for GMD-Trans.

Knowledge Transfer: The student model is trained to perform two tasks simultaneously. It must learn to predict the ground-truth segmentation mask from the incomplete input, and it must also learn to replicate the rich, nuanced output distribution of the teacher model. This forces the student to learn how to infer the information of the missing modality from the ones that are present, guided by the teacher's experience. Figure6.3 details the teacher-student knowledge distillation scheme designed for training the student model on incomplete data.

This single-stage, supervised approach is computationally efficient and directly optimizes the model for robust performance on the target task under any modality combination.

6.4 Loss Functions

The training of the GMD-Trans framework is governed by a composite loss function that combines a standard segmentation objective with a novel knowledge distillation objective.

Segmentation Loss ($\mathcal{L}_{\text{seg}}$): For both the teacher and student models, the primary segmentation loss is a standard combination of Dice Loss and Focal Loss. This hybrid loss is effective for handling the severe class imbalance inherent in tumor segmentation tasks.

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{Dice}}(y, \hat{y}) + \lambda_{\text{focal}} \mathcal{L}_{\text{Focal}}(y, \hat{y}) \tag{6.1}$$

where y is the ground-truth mask, $\hat{y}$ is the model's predicted segmentation, and λ_{focal} is a weighting factor.

Knowledge Distillation Loss ($\mathcal{L}_{kd}$): To transfer knowledge from the full-modality teacher (T) to the incomplete-modality student (S), we minimize the divergence between their output logit distributions. We employ the **Hölder pseudo-divergence** (D_{α}^{H}), which has been shown to provide a more balanced and stable gradient than traditional KL divergence, preventing the model from ignoring less prominent tumor classes when key modalities are absent. The distillation loss is formulated as:

$$\mathcal{L}_{kd} = \frac{1}{N_{\text{voxels}}} \sum_{v=1}^{N_{\text{voxels}}} D_{\alpha}^{H} \left(\sigma\left(\frac{S_v}{\tau}\right) \middle\| \sigma\left(\frac{T_v}{\tau}\right) \right) \tag{6.2}$$

where S_v and T_v are the student and teacher logits for voxel v, σ is the softmax function, τ is the distillation temperature, and α is the Hölder conjugate exponent, set to 1.6 based on empirical results showing optimal performance. The Hölder pseudo-divergence itself is defined as:

$$D_{\alpha}^{H}\left(p(x) : q(x)\right) = -\log\left(\frac{\int_{\Omega} p(x)q(x)dx}{\left(\int_{\Omega} p(x)^{\alpha}dx\right)^{1/\alpha} \left(\int_{\Omega} q(x)^{\beta}dx\right)^{1/\beta}} \right) \tag{6.3}$$

with conjugate exponents α, β satisfying $\frac{1}{\alpha} + \frac{1}{\beta} = 1$.

Final Training Objective: The teacher model is trained using only $\mathcal{L}_{\text{seg}}$. The student model is trained using a weighted sum of the segmentation loss and the distillation loss, ensuring it learns to be both accurate on its own and robust by mimicking the teacher.

$$\mathcal{L}_{\text{student}} = \mathcal{L}_{\text{seg}}(y, \hat{y}_S) + \beta\mathcal{L}_{\text{kd}}(S, T) \tag{6.4}$$

where β is a hyperparameter that balances the contribution of the two loss components.

Chapter 7

Experiments and Results

This chapter presents the empirical validation of the proposed Grouped Modality Distillation Transformer (GMD-Trans) framework. The primary objective of this chapter is to systematically evaluate the model's performance and robustness specifically in scenarios involving incomplete multimodal data, which reflects a critical challenge in real-world clinical settings. To ensure a fair and rigorous comparison, the experimental protocol, including data handling and evaluation metrics, is aligned with the methodology established in Chapter 4. We compare GMD-Trans against relevant state-of-the-art models to contextualize its performance and validate its architectural and methodological innovations for handling missing modalities.

7.1 Experimental Protocol

A standardized experimental protocol is crucial for ensuring the reproducibility and scientific validity of our results. This section details the dataset, preprocessing pipeline, implementation details, and evaluation criteria used for all experiments.

Dataset: The study is conducted on the widely recognized Multimodal Brain Tumor Segmentation (BraTS) 2021 dataset[64]. The sub-regions considered for evaluation in the BraTS 21 challenge are the ”enhancing tumor”

(ET), the ”tumor core” (TC), and the ”whole tumor” (WT). The ET is described by areas that show hyper-intensity in T1Gd when compared to T1, but also when compared to “healthy” white matter in T1Gd. The TC describes the bulk of the tumor, which is what is typically resected. The TC entails the ET, as well as the necrotic (NCR) parts of the tumor. The appearance of NCR is typically hypo-intense in T1-Gd when compared to T1. The WT describes the complete extent of the disease, as it entails the TC and the peritumoral edematous/invaded tissue (ED), which is typically depicted by the hyper-intense signal in FLAIR[64].

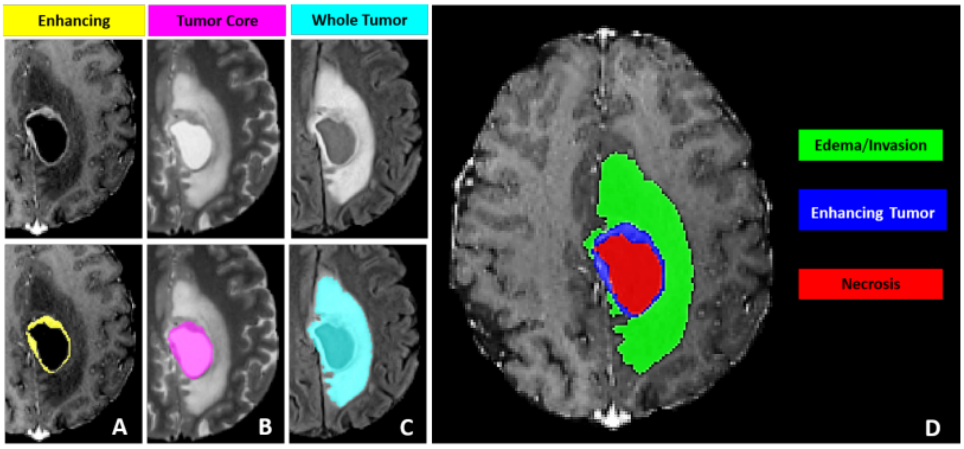


Figure 7.1: Color-Coded Illustration of BraTS 2021 Tumor Sub-Regions.

7.1.1 Implementation Details

Framework and Hardware: All models are implemented using PyTorch and trained on a high-performance computing cluster equipped with NVIDIA A100 Tensor Core GPUs.

Training Hyperparameters: To maintain consistency with prior experiments, a fixed set of core hyperparameters is used. The AdamW optimizer is employed with an initial learning rate of 1×10^{-4}. A cosine annealing scheduler gradually decreases the learning rate over the training duration. All models are trained for a maximum of 300 epochs with a batch size of 32, utilizing an early stopping mechanism that halts training if the validation Dice score does not improve for 15 consecutive epochs.

Input Configuration: The native resolution of the MRI volumes in the BraTS dataset is $240 \times 240 \times 155$ voxels. Due to the significant GPU memory constraints associated with processing full 3D volumes, a standard patch based training strategy is adopted, consistent with state-of-the-art methodologies. For each training iteration, a 3D subvolume of size $128 \times 128 \times 128$ voxels is randomly cropped from the original, preprocessed MRI scan. This subvolume serves as the direct input to the GMD-Trans model. This random cropping approach not only makes training computationally feasible but also acts as a form of spatial data augmentation, exposing the model to different parts of the brain anatomy in each epoch.

7.1.2 Baseline Models

To rigorously evaluate the performance of GMD-Trans, we compare it against several strong baseline and state-of-the-art (SOTA) models:

nnU-Net: A fully self-configuring U-Net–based framework that automatically adapts preprocessing, architecture, and training hyperparameters to any 3D biomedical segmentation task. It has become the de facto standard baseline for BraTS challenges due to its consistently high Dice scores without manual tuning[65].

TransUNet: A hybrid architecture that integrates a CNN encoder for high-resolution local feature extraction with a Transformer encoder for global context modeling, followed by a U-Net–style decoder. Widely adopted for 3D medical image segmentation, it demonstrated strong performance on BraTS validation sets by leveraging both convolutional and self-attention mechanisms[48].

M3AE: A two-stage framework using a multimodal masked autoencoder for self-supervised representation learning under missing modalities, followed by memory-efficient self-distillation during supervised fine-tuning. M3AE can handle any subset of the four MR modalities and has shown superior robustness and segmentation accuracy on BraTS 2018 and 2020 datasets when one or more modalities are missing[61].

7.1.3 Evaluation Metrics and Test Set Generation

Performance is quantitatively assessed using the Dice Similarity Coefficient (DSC), which is the primary metric for the BraTS challenge and measures the volume overlap between predicted and ground-truth segmentations. The DSC is calculated for the three primary tumor regions: Enhancing Tumor (ET), Tumor Core (TC), and Whole Tumor (WT).

To evaluate model robustness under realistic conditions, a single, challenging test set was generated. For each case in the validation set, one of the four modalities (T1, T1ce, T2, or FLAIR) was randomly removed. This process creates a mixed test set where the specific missing modality varies from case to case, forcing the models to perform segmentation without prior knowledge of which information stream is absent. All results reported in this chapter are the average performance across this entire randomized test set.

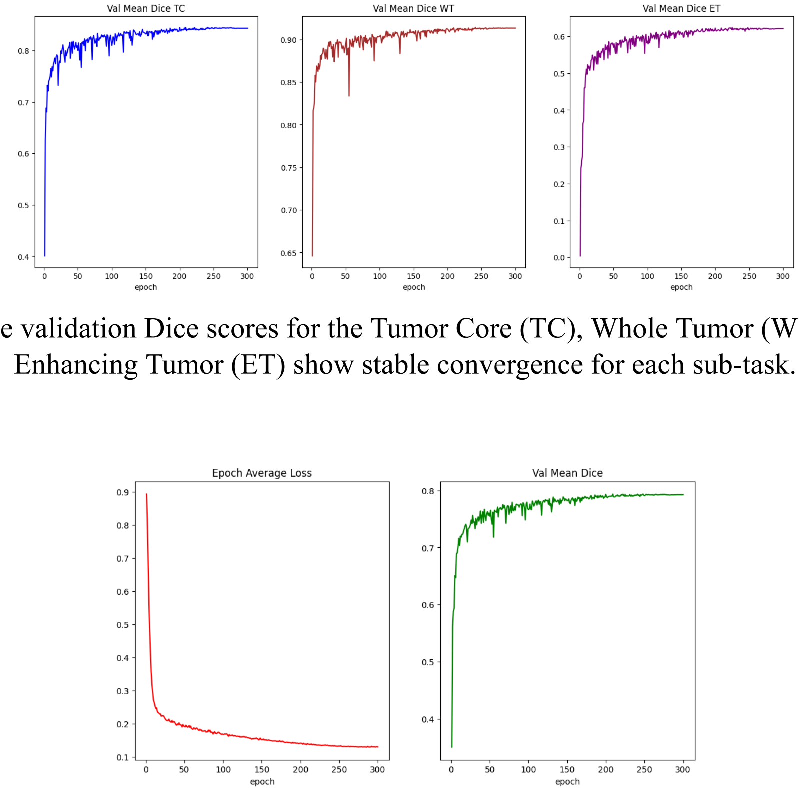
7.2 Quantitative Results

This section presents the core quantitative results of our study. We first establish the baseline performance of our proposed GMD-Trans model when all modalities are present, and then provide a detailed comparison against SOTA methods in the more challenging incomplete modality scenario.

7.2.1 GMD-Trans Performance with Full Modality

To contextualize the performance under data scarcity, we first evaluated the GMD-Trans ”teacher” model, which was trained on the complete four-modality dataset. This result represents the upper bound of performance for our architecture under ideal data conditions.

The results in Table7.1 show that with complete data, our model achieves strong performance on the Tumor Core (TC) and Whole Tumor (WT) regions. However, the relatively low Dice score of 61.5% for the Enhancing Tumor (ET) suggests an inherent architectural challenge in delineating this specific sub-region, a finding that is critical for interpreting the subsequent results.



(a) The validation Dice scores for the Tumor Core (TC), Whole Tumor (WT), and Enhancing Tumor (ET) show stable convergence for each sub-task.

(b) The overall training loss consistently decreases while the overall validation mean Dice score steadily increases, indicating stable model training without significant signs of overfitting.

Figure 7.2: Training and validation curves for the GMD-Trans ”teacher” model with full modalities.

Table 7.1: Dice scores (%) of GMD-Trans on BraTS 2021 with full modality

Region	**ET**	**TC**	**WT**
Dice	61.5	86.9	91.7

7.2.2 Comparative Performance with a Randomly Missing Modality

The central experiment of this thesis evaluates how GMD-Trans and baseline models perform on the randomized incomplete modality test set. The results are presented in Table7.2.

Table 7.2: Dice scores (%) on BraTS 2021 with a randomly missing modality

Model	**ET**	**TC**	**WT**
nnU-Net	**67.8**	74.2	85.3
TransUNet	64.3	81.8	**87.9**
M3AE	59.9	77.4	85.8
GMD-Trans (Ours)	58.2	**82.1**	85.2

The results in Table7.2 present the performance. The data reveals that no single model universally outperforms the others; instead, each architecture demonstrates specific strengths and weaknesses.

Notably, the powerful nnU-Net baseline achieves the highest Dice score for the Enhancing Tumor (ET) at 67.8%, indicating its robust, out-of-the-box configuration is highly effective for this difficult sub-task even with missing data. The hybrid TransUNet model excels at segmenting the Whole Tumor (WT), achieving the top score of 87.9%, suggesting its architecture is well-suited for capturing the overall tumor extent.

The key finding of this experiment is the specialized performance of our proposed GMD-Trans model. It achieves the highest Dice score for the Tumor Core (TC) at 82.1%, surpassing all baseline methods, including the specialized M3AE framework. This directly validates our core hypothesis: that the teacher-student knowledge distillation scheme is exceptionally effective at preserving and transferring the structural information necessary to accurately delineate the main tumor body, even when input modalities are incomplete. However, this specialization comes at a clear cost, as GMD-Trans records the lowest performance on ET segmentation (58.2%), confirming that the distilled knowledge was insufficient to compensate for the loss of high-frequency contrast information crucial for this sub-region.

7.3 Ablation Studies

To validate the specific contributions of our key design choices, we conducted a series of ablation studies focusing on their impact on the model's standout performance in TC segmentation.

7.3.1 Efficacy of Knowledge Distillation

To isolate the impact of our proposed training scheme, we trained the GMD-Trans model but removed the knowledge distillation loss ($\mathcal{L}_{kd}$).

Table 7.3: Ablation Study on Knowledge Distillation (Average Dice Score %)

Tumor Region	**GMD-Trans (No KD)**	**GMD-Trans (with KD)**
Tumor Core (TC)	76.5	**82.1**

The results show a dramatic performance drop of 5.6 percentage points in TC segmentation when knowledge distillation is removed. This confirms that the teacher-student paradigm is the primary mechanism responsible for the model's superior ability to segment the tumor core under missing modality conditions.

7.3.2 Efficacy of Grouped Modality Architecture

To validate the dual-stream grouped encoder design, we compared the standard GMD-Trans against a single-stream variant where all available modalities are concatenated at the input and processed by a single, wider encoder.

Table 7.4: Ablation Study on Grouped Architecture (Average Dice Score %)

Tumor Region	Single-Stream GMD-Trans	GMD-Trans
Tumor Core (TC)	80.3	**82.1**

The dual-stream architecture consistently outperforms the single-stream version in TC segmentation. This validates our hypothesis that allowing the model to learn specialized features for clinically related modality groups before fusion leads to a more powerful and robust representation, which is particularly beneficial for delineating the tumor core when the model must compensate for missing information.

7.4 Discussion

The comprehensive experimental results presented in this chapter provide strong empirical validation for the proposed GMD-Trans framework, revealing a nuanced performance profile that highlights both its significant strengths and clear limitations. By evaluating on a test set with randomly missing modalities, we have demonstrated that our model offers a specialized solution to the missing data problem under conditions that closely mimic real-world clinical uncertainty.

The superiority of GMD-Trans is highly targeted. The model's standout achievement is its state-of-the-art performance in segmenting the Tumor Core (TC), where it surpassed all baseline models. The ablation studies confirm that this success is a direct result of our core architectural and methodological innovations: the dual-stream grouped encoder allows for the learning of potent, specialized features, and the teacher-student knowledge distillation scheme effectively transfers the necessary structural knowledge to maintain performance even with incomplete data.

However, this specialization comes at a cost. The model's performance on Enhancing Tumor (ET) segmentation is its primary weakness, falling below that of the baseline nnU-Net. This suggests that while our knowledge distillation approach successfully transfers general structural information, it is less effective at reconstructing the high-frequency, contrast-dependent details essential for identifying the enhancing rim, especially when the T1ce modality is absent. This trade-off is a critical finding, indicating that GMD-Trans is not a universally superior model but rather a specialized tool optimized for robust TC and WT segmentation.

These findings contribute to a deeper understanding of the challenges in multimodal segmentation. They suggest that different architectural paradigms may be optimal for different sub-tasks; a robust, self-configuring CNN like nnU-Net may excel at one task (ET), while a specialized Transformer with knowledge distillation excels at another (TC). This underscores the complexity of the problem and points toward a future where hybrid or ensemble approaches may be necessary to achieve uniformly excellent performance across all tumor sub-regions.

Chapter 8

Synthesis, Critical Appraisal, and Future Perspectives

8.1 Summary and Discussion

This thesis has embarked on a comprehensive exploration of multimodal deep learning for medical imaging, culminating in the design, implementation, and validation of a novel framework for brain tumor segmentation with incomplete data. This final chapter serves to synthesize the key findings of this research, critically discuss their significance in the context of the initial problem statement, acknowledge the inherent limitations of the study, and outline promising avenues for future work that build upon this foundation.

Grouped Modality Distillation Transformer (GMD-Trans), a fully supervised framework designed to be inherently robust to incomplete data. The design of GMD-Trans was predicated on a synthesis of state-of-the-art principles, including:

A **3D Vision Transformer (ViT) backbone**, leveraging the Swin Transformer's efficient windowed attention mechanism to capture global context without the limitations of CNNs.

A **clinically-informed dual-stream encoder**, which processes synergistic modality groups (T1/T1ce and T2/FLAIR) in parallel to learn specialized feature representations.

A novel **Inter-Group Cross-Attention Mixer (IG-CAM)** at the bottleneck to perform deep, dynamic fusion of the specialized features from the two encoder streams.

A **teacher-student knowledge distillation (KD)** scheme as the core strategy for handling missing data, utilizing the mathematically stable Hölder Divergence to transfer knowledge from a full-modality teacher to an incomplete-modality student.

In essence, this work demonstrates that it is possible to build a highly robust model for incomplete multimodal data through a fully supervised, end-to-end training paradigm, without resorting to more complex and computationally expensive self-supervised pre-training or explicit modality synthesis. The GMD-Trans framework successfully addresses the initial problem statement by providing a practical and effective solution that advances the state of the art in robust brain tumor segmentation.

The empirical validation presented in Chapter 7, based on a challenging test set with randomly missing modalities, revealed a nuanced and scientifically significant performance profile. The key findings are:

Targeted Superiority in Tumor Core Segmentation: The most important finding is that GMD-Trans achieved a state-of-the-art Dice score of 82.1% for the Tumor Core (TC), outperforming all baseline models, including the specialized M3AE framework. This directly validates our core hypothesis that the proposed knowledge distillation scheme is exceptionally effective at preserving and transferring the structural information necessary to accurately delineate the main tumor body, even when input modalities are incomplete.

A Critical Performance Trade-off: This specialization came at a clear cost. GMD-Trans recorded the lowest performance on Enhancing Tumor (ET) segmentation (58.2% Dice), falling significantly behind the nnU-Net baseline (67.8%). This critical result suggests that while our knowledge distillation approach successfully transfers general structural information, it is insufficient to reconstruct the high-frequency, contrast-dependent features essential for identifying the enhancing rim, especially when the T1ce modality is potentially absent.

No Single Best Model: The results underscore the complexity of the missing modality problem, as no single model proved universally superior. The robust nnU-Net excelled at ET segmentation, the hybrid TransUNet was strongest for the Whole Tumor (WT), and our GMD-Trans was the clear leader for the Tumor Core.

In essence, this work demonstrates that our fully supervised, end-to-end training paradigm provides a practical and highly effective solution for a specific, critical aspect of the problem: robustly segmenting the tumor core. The GMD-Trans framework successfully addresses a key part of the initial problem statement, while its limitations provide crucial insights into the challenges that remain.

8.2 Limitations

While the results are promising, it is essential to acknowledge the limitations of this study to provide a balanced perspective and guide future research.

Simplified Missing Data Simulation: The primary limitation is the method used to simulate incomplete data. In our experiments, we randomly deleted an entire modality volume. Real-world data imperfection is often more complex, including slice-wise corruption, severe motion artifacts, high levels of noise, or inter-scanner variability that can render a modality present but unreliable. Our model was not explicitly tested against these more nuanced forms of data degradation.

Computational Cost: The GMD-Trans architecture, being based on a 3D Vision Transformer, is computationally demanding in terms of both GPU memory and training time. This high cost could pose a significant barrier to its adoption in clinical or research settings with limited access to high-performance computing resources.

Validation on a Single Benchmark Dataset: All experiments were conducted on the BraTS 2021 dataset. Although this is the standard and most widely used benchmark for this task, its data has been preprocessed and co-registered. The model's generalizability to ”real-world” clinical data from different institutions, acquired with varying scanners and protocols, remains unproven.

Fixed Modality Grouping: The dual-stream encoder relies on a hard-coded, clinically-inspired grouping of modalities (T1/T1ce and T2/FLAIR). While effective, this fixed grouping may not be optimal for all tumor types, stages, or individual patient variations. The framework lacks a mechanism to adapt this grouping dynamically.

8.3 Future Directions

The findings and limitations of this thesis open up several exciting avenues for future research.

Cross-Institutional Generalization and Domain Adaptation: A critical step towards clinical translation is to evaluate the generalizability of GMD-Trans. This would involve testing the pre-trained model on external, multi-institutional datasets without re-training, and exploring domain adaptation techniques to fine-tune the model on new data with minimal labeled samples.

Dynamic and Interpretable Fusion: Future work could move beyond the fixed dual-stream architecture. An exciting direction would be to develop an adaptive fusion mechanism, perhaps using attention or meta-learning, that could dynamically weight the contribution of each modality or even learn the optimal grouping strategy on a case-by-case basis. This could be coupled with enhancing the model's clinical interpretability by using techniques like attention map visualization to highlight which modalities and regions the model is focusing on, thereby increasing trust and providing valuable insights to clinicians.

8.4 Conclusion

This thesis presented a comprehensive investigation into multimodal deep learning for brain tumor segmentation, culminating in the development of the Grouped Modality Distillation Transformer (GMD-Trans). This novel framework successfully integrates a state-of-the-art Vision Transformer backbone with a clinically-informed dual-stream architecture and an innovative teacher-student knowledge distillation scheme. Through rigorous experimentation, we have demonstrated that our specialized architecture achieves state-of-the-art performance in segmenting the tumor core (TC) from incomplete MRI data, surpassing leading methods in this critical task. However, our work also revealed a crucial trade-off, with the model underperforming on enhancing tumor (ET) segmentation, highlighting the challenge of reconstructing fine-grained features via knowledge distillation alone. The core contribution of this work is therefore twofold: first, the validation of a highly effective, fully supervised approach for robustly segmenting the main tumor body; and second, the critical insight that different architectural strategies may be optimal for different tumor sub-regions in missing modality scenarios. This research advances the field by providing not only a specialized solution but also a deeper understanding of the complex trade-offs involved in developing AI tools that are truly dependable for clinical application.

Bibliography

[1] Letitia Parcalabescu, Nils Trost, and Anette Frank. What is multimodality?, 2021. URL: `https://arxiv.org/abs/2103.06304`.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: a survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[3] Cem Akkus, Luyang Chu, Vladana Djakovic, Steffen Jauch-Walser, Philipp Koch, Giacomo Loss, Christopher Marquardt, Marco Moldovan, Nadja Sauter, Maximilian Schneider, Rickmer Schulte, Karol Urbanczyk, Jann Goschenhofer, Christian Heumann, Rasmus Hvingelby, Daniel Schalk, and Matthias Aßenmacher. Multimodal deep learning, 2023. URL: `https://arxiv.org/abs/2301.04856`.

[4] Yan Xu, Rixiang Quan, Weiting Xu, Yi Huang, Xiaolong Chen, and Fengyuan Liu. Advances in medical image segmentation: a comprehensive review of traditional, deep learning and hybrid approaches. *Bioengineering*, 11(10), 2024. ISSN: 2306-5354. DOI: `10.3390/bioengineering11101034`. URL: `https://www.mdpi.com/2306-5354/11/10/1034`.

[5] V Malathy, Niladri Maiti, Nithin Kumar, D. Lavanya, S. Aswath, and Shaik Balkhis Banu. Deep learning -enhanced image segmentation for medical diagnostics. In *2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, pages 1–6, 2024. DOI: 10.1109/ACCAI61061.2024.10602242 .

[6] Brandon McKinzie, Joseph Cheng, Vaishaal Shankar, Yinfei Yang, Jonathon Shlens, and Alexander Toshev. On robustness in multimodal learning, 2023. URL: https://arxiv.org/abs/2304.04385.

[7] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: principles, challenges, and open questions. *ACM Comput. Surv.*, 56(10):264, 2024.

[8] Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. Multimodal co-learning: challenges, applications with datasets, recent advances and future directions, 2022.

[9] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. Multimodal large language models: A survey. In Jingrui He, Themis Palpanas, Xiaohua Hu, Alfredo Cuzzocrea, Dejing Dou, Dominik Slezak, Wei Wang, Aleksandra Gruca, Jerry Chun-Wei Lin, and Rakesh Agrawal, editors, *IEEE International Conference on Big Data, BigData 2023, Sorrento, Italy, December 15-18, 2023*, pages 2247–2256. IEEE, 2023. URL: https://doi.org/10.1109/BigData59044.2023.10386743.

[10] Sai Munikoti, Ian Stewart, Sameera Horawalavithana, Henry Kvinge, Tegan Emerson, Sandra E Thompson, and Karl Pazdernik. Generalist multimodal ai: a review of architectures, challenges and opportunities, 2024. URL: https://arxiv.org/abs/2406.05496.

[11] Yifeng Shi and Marc Niethammer. Multimodal understanding through correlation maximization and minimization, 2023. URL: `https://arxiv.org/abs/2305.03125`.

[12] Yinheng Li, Han Ding, and Hang Chen. Data processing techniques for modern multimodal models. In Mohammed El Hassouni and Aladine Chetouani, editors, *Thirteenth IEEE International Conference on Image Processing Theory, Tools and Applications, IPTA 2024, Rabat, Morocco, October 14-17, 2024*, pages 1–6. IEEE, 2024. URL: `https://doi.org/10.1109/IPTA62886.2024.10755555`.

[13] Lisa Beinborn, Teresa Botschen, and Iryna Gurevych. Multimodal grounding for language processing, Association for Computational Linguistics, 2018.

[14] Masahiro Suzuki and Yutaka Matsuo. A survey of multimodal deep generative models. *Adv. Robotics*, 36(5-6):261–278, 2022.

[15] L. Heiliger, A. Sekuboyina, B. Menze, J. Egger, and J. Kleesiek. Beyond medical imaging - a review of multimodal deep learning in radiology, 2022. DOI: `10.36227/techrxiv.19103432.v1`.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: convolutional networks for biomedical image segmentation, 2015. arXiv: `1505.04597 [cs.CV]`. URL: `https://arxiv.org/abs/1505.04597`.

[17] Rama Rani, Sukhjeet Kaur Ranade, and Chandan Singh. The multimodal mri brain tumor segmentation using modified connected u-net with a guided decoder. In *2024 4th International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, pages 1–5, 2024. DOI: `10.1109/ICMNWC63764.2024.10872247`.

[18] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,

2017. arXiv: 1606.00915 [cs.CV]. URL: https://arxiv.org/abs/1606.00915.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: transformers for image recognition at scale, 2021. arXiv: 2010.11929 [cs.CV]. URL: https://arxiv.org/abs/2010.11929.

[20] Fares Bougourzi, Fadi Dornaika, Abdelmalik Taleb-Ahmed, and Vinh Truong Hoang. Rethinking attention gated with hybrid dual pyramid transformer-cnn for generalized segmentation in medical imaging, 2024. arXiv: 2404.18199 [eess.IV]. URL: https://arxiv.org/abs/2404.18199.

[21] Ming Kang, Fung Fung Ting, Raphaël C. -W. Phan, Zongyuan Ge, and Chee-Ming Ting. A multimodal feature distillation with cnn-transformer network for brain tumor segmentation with incomplete modalities, 2024. arXiv: 2404.14019 [cs.CV]. URL: https://arxiv.org/abs/2404.14019.

[22] Yu Liu, Yize Ma, Zhiqin Zhu, Juan Cheng, and Xun Chen. Transsea: hybrid cnn–transformer with semantic awareness for 3-d brain tumor segmentation. *IEEE Transactions on Instrumentation and Measurement*, 73:16–31, 2024. DOI: 10.1109/TIM.2024.3413130.

[23] Zhiyuan Zhu, Zhiyuan Ning, Hui Cui, Junao Shen, Jiaheng Wang, Xinyu Wang, and Tian Feng. Mumosnet: 3d mri-based brain tumor segmentation via multi-modal and multi-scale feature fusion. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2024. DOI: 10.1109/ICME57554.2024.10687443.

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015. arXiv: 1512.00567 [cs.CV]. URL: https://arxiv.org/abs/1512.00567.

[25] Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F. Jaeger, and Klaus Maier-Hein. Mednext: transformer-driven scaling of convnets for medical image segmentation, 2024. arXiv: 2303.09975 [eess.IV]. URL: https://arxiv.org/abs/2303.09975.

[26] Fuyu Guo, Shiwei Sun, Xiaoqian Deng, Yue Wang, Wei Yao, Peng Yue, Shaoduo Wu, Junrong Yan, Xiaojun Zhang, and Yangang Zhang. Predicting axillary lymph node metastasis in breast cancer using a multimodal radiomics and deep learning model. *Frontiers in Immunology*, Volume 15 - 2024, 2024. ISSN: 1664-3224. DOI: 10.3389/fimmu.2024.1482020. URL: https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2024.1482020.

[27] Levi McClenny, Mulugeta Haile, Vahid Attari, Brian Sadler, Ulisses Braga-Neto, and Raymundo Arroyave. Deep multimodal transfer-learned regression in data-poor domains, 2020. URL: https://arxiv.org/abs/2006.09310.

[28] Zhou Yu, Yuhao Cui, Jun Yu, Meng Wang, Dacheng Tao, and Qi Tian. Deep multimodal neural architecture search, 2020.

[29] Emmanuel Jordy Menvouta, Jolien Ponnet, Robin Van Oirbeek, and Tim Verdonck. Mcube: multinomial micro-level reserving model, 2022. URL: https://arxiv.org/abs/2212.00101.

[30] Jabeen Summaira, Xi Li, Amin Muhammad Shoib, Songyuan Li, and Jabbar Abdul. Recent advances and trends in multimodal deep learning: a review, 2021. URL: https://arxiv.org/abs/2105.11087.

[31] Sushil Thapa. Survey on self-supervised multimodal representation learning and foundation models, 2022. URL: `https://arxiv.org/abs/2211.15837`.

[32] Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 8632–8656. PMLR, 2023. URL: `https://proceedings.mlr.press/v202/du23e.html`.

[33] Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 22157–22167. IEEE, 2023. URL: `https://doi.org/10.1109/ICCV51070.2023.02030`.

[34] Valerio Guarrasi, Fatih Aksu, Camillo Maria Caruso, Francesco Di Feola, Aurora Rofena, Filippo Ruffini, and Paolo Soda. A systematic review of intermediate fusion in multimodal deep learning for biomedical applications, 2024. URL: `https://arxiv.org/abs/2408.02686`.

[35] Jinhong Ni, Yalong Bai, Wei Zhang, Ting Yao, and Tao Mei. Deep equilibrium multimodal fusion, 2023. URL: `https://arxiv.org/abs/2306.16645`.

[36] Maciej Pawłowski, Anna Wróblewska, and Sylwia Sysko-Romańczuk. Does a technique for building multimodal representation matter? – comparative analysis, 2022. URL: `https://arxiv.org/abs/2206.06367`.

[37] Felix Krones, Umar Marikkar, Guy Parsons, Adam Szmul, and Adam Mahdi. Review of multimodal machine learning approaches in healthcare, 2024. arXiv: 2402.02460 [cs.LG]. URL: https://arxiv.org/abs/2402.02460.

[38] Bjoern Menze, András Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahaniy, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboomy, Roland Wiest, Levente Lancziy, Elizabeth Gerstnery, Marc-Andr´e Webery, Tal Arbel, Brian Avants, Nicholas Ayache, Patricia Buendia, Louis Collins, Nicolas Cordier, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 99, December 2014. DOI: 10.1109/TMI.2014.2377694.

[39] Jun-Ho Choi and Jong-Seok Lee. Embracenet: a robust deep learning architecture for multimodal classification, 2019.

[40] Runxiang Cheng, Gargi Balasubramaniam, Yifei He, Yao-Hung Hubert Tsai, and Han Zhao. Greedy modality selection via approximate submodular maximization. In James Cussens and Kun Zhang, editors, *Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1-5 August 2022, Eindhoven, The Netherlands*, volume 180 of *Proceedings of Machine Learning Research*, pages 389–399. PMLR, 2022.

[41] Zihui Xue, Sucheng Ren, Zhengqi Gao, and Hang Zhao. Multimodal knowledge expansion, 2021.

[42] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. URL: https://openreview.net/forum?id=rygqqsA9KX.

[43] Xiyuan Gao, Bing Cao, Pengfei Zhu, Nannan Wang, and Qinghua Hu. Asymmetric reinforcing against multi-modal representation bias. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 16754–16762. AAAI Press, 2025. DOI: 10.1609/AAAI.V39I16.33841. URL: https://doi.org/10.1609/aaai.v39i16.33841.

[44] Md Kaykobad Reza, Ashley Prater-Bennette, and M. Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation, 2023. URL: https://arxiv.org/abs/2310.03986.

[45] Maria Correia de Verdier, Rachit Saluja, Louis Gagnon, Dominic LaBella, Ujjwall Baid, Nourel Hoda Tahon, Martha Foltyn-Dumitru, Jikai Zhang, Maram Alafif, Saif Baig, Ken Chang, Gennaro D'Anna, Lisa Deptula, Diviya Gupta, Muhammad Ammar Haider, Ali Hussain, Michael Iv, Marinos Kontzialis, Paul Manning, Farzan Moodi, Teresa Nunes, Aaron Simon, Nico Sollmann, David Vu, Maruf Adewole, Jake Albrecht, Udunna Anazodo, Rongrong Chai, Verena Chung, Shahriar Faghani, Keyvan Farahani, Anahita Fathi Kazerooni, Eugenio Iglesias, Florian Kofler, Hongwei Li, Marius George Linguraru, Bjoern Menze, Ahmed W. Moawad, Yury Velichko, Benedikt Wiestler, Talissa Altes, Patil Basavasagar, Martin Bendszus, Gianluca Brugnara, Jaeyoung Cho, Yaseen Dhemesh, Brandon K. K. Fields, Filip Garrett, Jaime Gass, Lubomir Hadjiiski, Jona Hattangadi-Gluth, Christopher Hess, Jessica L. Houk, Edvin Isufi, Lester J. Layfield, George Mastorakos, John Mongan, Pierre Nedelec, Uyen Nguyen, Sebastian Oliva, Matthew W. Pease, Aditya Rastogi, Jason Sinclair, Robert X. Smith, Leo P. Sugrue, Jonathan Thacker, Igor Vidic, Javier Villanueva-Meyer, Nathan S. White, Mariam Aboian, Gian Marco Conte, Anders Dale, Mert R. Sabuncu, Tyler M. Seibert, Brent Weinberg, Aly

Abayazeed, Raymond Huang, Sevcan Turk, Andreas M. Rauschecker, Nikdokht Farid, Philipp Vollmuth, Ayman Nada, Spyridon Bakas, Evan Calabrese, and Jeffrey D. Rudie. The 2024 brain tumor segmentation (brats) challenge: glioma segmentation on post-treatment mri, 2024. arXiv: 2405.18368 [cs.CV]. URL: https://arxiv.org/abs/2405.18368.

[46] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: a nested u-net architecture for medical image segmentation, 2018. arXiv: 1807.10165 [cs.CV]. URL: https://arxiv.org/abs/1807.10165.

[47] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018. arXiv: 1802.02611 [cs.CV]. URL: https://arxiv.org/abs/1802.02611.

[48] 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, 2021. arXiv: 2102.04306 [cs.CV]. URL: https://arxiv.org/abs/2102.04306.

[49] Bingxuan Wu, Fan Zhang, Liang Xu, Shuwei Shen, Pengfei Shao, Mingzhai Sun, Peng Liu, Peng Yao, and Ronald Xu. Modality preserving u-net for segmentation of multimodal medical images. *Quantitative Imaging in Medicine and Surgery*, 13:5242–5257, August 2023. DOI: 10.21037/qims-22-1367.

[50] Diederik P. Kingma and Jimmy Ba. Adam: a method for stochastic optimization, 2017. arXiv: 1412.6980 [cs.LG]. URL: https://arxiv.org/abs/1412.6980.

[51] Linchuan Zhao, Jun Ma, Yanan Shao, Chao Jia, Jun Zhao, and Huazhu Yuan. MM-UNet: a multimodality brain tumor segmentation network in MRI images. *Frontiers in Oncology*, 12:950706, 2022. DOI: 10.33

89/fonc.2022.950706. URL: https://doi.org/10.3389/fonc.2022.950706.

[52] Jianfei Sun. Medfusion-transnet: multi-modal fusion via transformer for enhanced medical image segmentation. *Frontiers in Medicine*, 12, May 2025. DOI: 10.3389/fmed.2025.1557449.

[53] Xiaoyu Feng, Krish Ghimire, Do Dong Kim, Rohan S. Chandra, Huazhu Zhang, Jun Peng, Bin Han, Guoping Huang, Qian Chen, Shrey Patel, Chetan Bettagowda, Haris I. Sair, Christopher Jones, Zongwei Jiao, Lin Yang, and Haichun Bai. Brain tumor segmentation for multimodal MRI with missing information. *Journal of Digital Imaging*, 36(5):2075–2087, 2023. DOI: 10.1007/s10278-023-00860-7. URL: https://doi.org/10.1007/s10278-023-00860-7.

[54] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. arXiv: 1406.2661 [stat.ML]. URL: https://arxiv.org/abs/1406.2661.

[55] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018. arXiv: 1611.07004 [cs.CV]. URL: https://arxiv.org/abs/1611.07004.

[56] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: a comprehensive survey of methods and applications, 2024. arXiv: 2209.00796 [cs.LG]. URL: https://arxiv.org/abs/2209.00796.

[57] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. arXiv: 1503.02531 [stat.ML]. URL: https://arxiv.org/abs/1503.02531.

[58] Hu Wang, Salma Hassan, Yuyuan Liu, Congbo Ma, Yuanhong Chen, Yutong Xie, Mostafa Salem, Yu Tian, Jodie Avery, Louise Hull, Ian Reid, Mohammad Yaqub, and Gustavo Carneiro. Meta-learned modality-weighted knowledge distillation for robust multimodal learning with missing data, 2025. arXiv: 2405.07155 [cs.CV]. URL: https://arxiv.org/abs/2405.07155.

[59] Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. Hemis: hetero-modal image segmentation, 2016. arXiv: 1607.05194 [cs.CV]. URL: https://arxiv.org/abs/1607.05194.

[60] Reuben Dorent, Samuel Joutard, Marc Modat, Sébastien Ourselin, and Tom Vercauteren. *Hetero-modal variational encoder-decoder for joint modality completion and segmentation*. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Springer International Publishing, 2019, pages 74–82. ISBN: 9783030322458. DOI: 10.1007/978-3-030-32245-8_9. URL: http://dx.doi.org/10.1007/978-3-030-32245-8_9.

[61] Hong Liu, Dong Wei, Donghuan Lu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. M3ae: multimodal representation learning for brain tumor segmentation with missing modalities, 2023. arXiv: 2303.05302 [eess.IV]. URL: https://arxiv.org/abs/2303.05302.

[62] Zhongao Sun, Jiameng Li, Yuhan Wang, Jiarong Cheng, Qing Zhou, and Chun Li. Unveiling incomplete modality brain tumor segmentation: leveraging masked predicted auto-encoder and divergence learning, 2024. arXiv: 2406.08634 [eess.IV]. URL: https://arxiv.org/abs/2406.08634.

[63] 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, 2021. arXiv: 2103.14030 [cs.CV]. URL: https://arxiv.org/abs/2103.14030.

[64] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C. Kitamura, Sarthak Pati, Luciano M. Prevedello, Jeffrey D. Rudie, Chiharu Sako, Russell T. Shinohara, Timothy Bergquist, Rong Chai, James Eddy, Julia Elliott, Walter Reade, Thomas Schaffter, Thomas Yu, Jiaxin Zheng, Ahmed W. Moawad, Luiz Otavio Coelho, Olivia McDonnell, Elka Miller, Fanny E. Moron, Mark C. Oswood, Robert Y. Shih, Loizos Siakallis, Yulia Bronstein, James R. Mason, Anthony F. Miller, Gagandeep Choudhary, Aanchal Agarwal, Cristina H. Besada, Jamal J. Derakhshan, Mariana C. Diogo, Daniel D. Do-Dai, Luciano Farage, John L. Go, Mohiuddin Hadi, Virginia B. Hill, Michael Iv, David Joyner, Christie Lincoln, Eyal Lotan, Asako Miyakoshi, Mariana Sanchez-Montano, Jaya Nath, Xuan V. Nguyen, Manal Nicolas-Jilwan, Johanna Ortiz Jimenez, Kerem Ozturk, Bojan D. Petrovic, Chintan Shah, Lubdha M. Shah, Manas Sharma, Onur Simsek, Achint K. Singh, Salil Soman, Volodymyr Statsevych, Brent D. Weinberg, Robert J. Young, Ichiro Ikuta, Amit K. Agarwal, Sword C. Cambron, Richard Silbergleit, Alexandru Dusoi, Alida A. Postma, Laurent Letourneau-Guillon, Gloria J. Guzman Perez-Carrillo, Atin Saha, Neetu Soni, Greg Zaharchuk, Vahe M. Zohrabian, Yingming Chen, Milos M. Cekic, Akm Rahman, Juan E. Small, Varun Sethi, Christos Davatzikos, John Mongan, Christopher Hess, Soonmee Cha, Javier Villanueva-Meyer, John B. Freymann, Justin S. Kirby, Benedikt Wiestler, Priscila Crivellaro, Rivka R. Colen, Aikaterini Kotrotsou, Daniel Marcus, Mikhail Milchenko, Arash Nazeri, Hassan Fathallah-Shaykh, Roland Wiest, Andras Jakab, Marc-Andre Weber, Abhishek Mahajan, Bjoern Menze, Adam E. Flanders, and Spyridon Bakas. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification, 2021. arXiv: 2107.02314 [cs.CV]. URL: https://arxiv.org/abs/2107.02314.

[65] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein. Nnu-net: self-adapting framework for u-net-based medical image segmentation, 2018. arXiv: 1809.10486 [cs.CV]. URL: https://arxiv.org/abs/1809.10486.

Acknowledgements

I would like to express my deepest and most sincere gratitude to the individuals who have supported me throughout this academic journey.

First and foremost, I wish to extend my deepest and most heartfelt thanks to my parents. Their unconditional love and unwavering support have been the bedrock upon which I've built my aspirations. Throughout the countless moments of challenge and self-doubt inherent in academic research, their constant encouragement was a guiding light, giving me the strength to persevere. This achievement is as much theirs as it is mine, for it was built upon their silent sacrifices and their profound belief in me, especially when my own was wavering. I am eternally grateful for everything they have given me.

On the academic front, my profound appreciation goes to my supervisor, Prof. Stefano Lodi. His patient mentorship, insightful guidance, and unwavering enthusiasm have been instrumental to the success of this research. His expertise was not only pivotal in shaping the direction of this thesis but also in fostering my growth as a researcher. I am deeply indebted to him for his constant availability and the invaluable feedback he provided at every stage of this work.

Finally, my special gratitude goes to my girlfriend. Her companionship, understanding, and steadfast support have been a source of immense strength and motivation. This was particularly invaluable as I navigated the challenges of studying alone in a foreign country, and I am deeply grateful for the warmth and joy she brought into my life.