Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

EMSA-Net: Enhancing Brain Tumor Segmentation with Equivalent Multi-Scale and Cross-Modality Attention under limited MRI Modalities

Beibei Hou
Henan Polytechnic University
Yamin Wang
wangyamin@home.hpu.edu.cn

Henan Polytechnic University
Zhaozhao Xu
Henan Polytechnic University
Junding Sun
Henan Polytechnic University
Lihong Zhang
Henan Polytechnic University
Lei Zhao
Hunan University
Bin Pu
Hong Kong University of Science and Technology
Yudong Zhang
Henan Polytechnic University

Research Article

Keywords: Limited modality, brain tumor segmentation, feature fusion, cross-modality, knowledge distillation

Posted Date: February 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8701857/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

EMSA-Net: Enhancing Brain Tumor Segmentation with Equivalent Multi-Scale and Cross-Modality Attention under limited MRI Modalities

Beibei Hou[a], Yamin Wang[a], Zhaozhao Xu[a], Junding Sun[a], Lihong Zhang[a], Lei Zhao[b], Bin Pu[c], Yudong Zhang[a,d,e,*]

[a]*School of Computer Science and Technology, Henan Polytechnic University, Century Road 2001, Jiaozuo, 454000, Henan, China*
[b]*College of Computer Science and Electronic Engineering, Hunan University, Changsha, 410082, Hunan, China*
[c]*Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China*
[d]*School of Computing and Mathematics, University of Leicester, Leicester, LE1 7RH, UK*
[e]*School of Computer Science and Engineering, Southeast University, Nanjing, 210096, Jiangsu, China*

Abstract

Multi-modality medical imaging is essential for disease diagnosis and clinical treatment as it can provide complementary information on the morphological features of tumors. Clinically, acquiring four complete modalities is difficult due to device limitations, scan time, and limited costs, such that getting high segmentation performance for brain tumors becomes more challenging. To address the challenges posed by limited MRI modalities in clinical scenarios, the network that contains two innovations: Equivalent Multi-Scale Module and Cross-Modality Attention Mechanism (EMSA-Net) is proposed to exploit the effective features of limited modality to enhance feature representation and increase clinical applicability. Remarkably, Equivalent Multi-

*Corresponding author

Email addresses: houbeibei0120@hpu.edu.cn (Beibei Hou), wangyamin@home.hpu.edu.cn (Yamin Wang), xzz@hpu.edu.cn (Zhaozhao Xu), sunjd@hpu.edu.cn (Junding Sun), lihongzhang@hup.edu.cn (Lihong Zhang), zhaolei@hnu.edu.cn (Lei Zhao), eebinpu@ust.hk (Bin Pu), yudongzhang@ieee.org (Yudong Zhang)

Scale Module (EMS) utilizes the available modalities to extract brain tumor regions fully and integrates features according to the contributions of available modalities. Cross-Modality Attention Mechanism (CMA) further fuses multi-modality features and mitigates the impact of limited modalities in medical image segmentation. In this way, EMSA-Net can effectively capture the relationship between different modalities and tumor regions, enhancing the robustness of the model. Additionally, this paper introduces a knowledge distillation method where the student network is trained on a subset of the teacher's inputs. This approach allows the student model to learn from the teacher model's knowledge, enabling effective segmentation even with limited modalities. The proposed method is evaluated on the BraTS2020 dataset, demonstrating its effectiveness in improving segmentation performance under limited modality conditions. The performance of the proposed model is superior to state-of-the-art methods. For all 15 multi-modality combinations, the proposed method achieves average Dice scores of 88.7% for WT, 82.7% for TC, and 68.8% for ET, outperforming other methods by 0.8%, 2.9%, and 2.1% respectively. More important, limited modalities, such as T1ce, also perform comparably to complete modalities.

Keywords: Limited modality, brain tumor segmentation, feature fusion, cross-modality, knowledge distillation

---

1. Introduction

Accurate and automated segmentation of brain tumors plays a pivotal role in diagnosis, treatment planning, and prognosis evaluation [1, 2]. Multi-modality imaging is widely used in medical image analysis because it provides complementary biological information for brain tumor segmentation, thereby improving both accuracy and reliability. However, acquiring a complete set of multi-modality MRI data poses significant challenges. Limitations arise from device constraints, prolonged scanning time, and high costs, which can result in incomplete modality availability in clinical practice. As a result, developing a robust medical image segmentation method to address the limited availability modality challenge has become an urgent and essential research focus.

In particular, Magnetic Resonance Imaging (MRI) [3] with its multiple pulse sequences, enables the generation of diverse tissue contrasts that help identify anatomical structures and detect functional tissue changes. In detail,

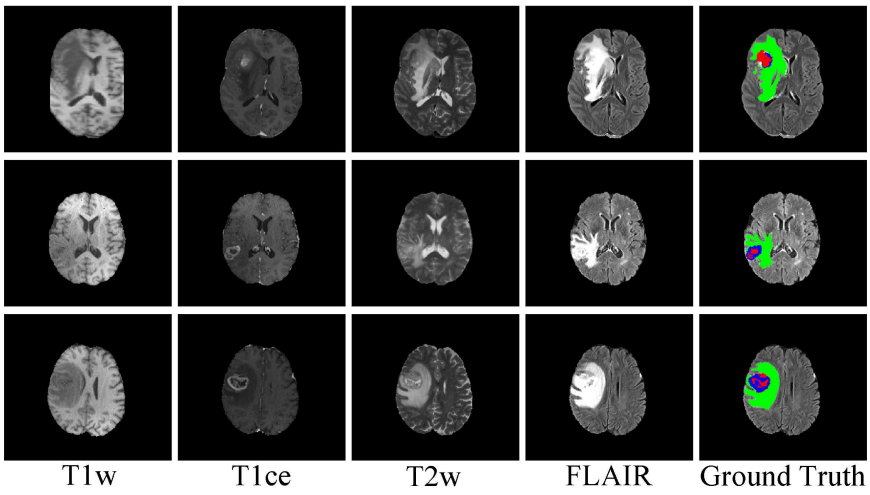


Figure 1: Visualization of brain tumor segmentation across multiple MRI modalities. From left to right: MRI images in four modalities (T1w, T1ce, T2w, and FLAIR) are represented for three different patients. The corresponding ground truth labels for each patient are shown, with distinct colors indicating different tumor regions in the segmentation results.

various MRI modalities, include T1-weighted (T1w), contrast-enhanced T1-weighted (T1ce), T2-weighted (T2w), and Fluid Attenuation Inversion Recovery (FLAIR). Specifically, various modalities exhibit unique visual characteristics and respond differently to diverse tumor regions. For example, as shown in Figure 1, T1w is primarily used to observe brain tissue structures, T1ce highlights the tumor core and enhanced regions in blue and red colors, T2w emphasizes the extent of lesions and FLAIR focuses on edema and infiltrative abnormalities in green color. These distinct responses to tumor regions highlight the importance of modality-specific information in brain tumor segmentation.

Recently, deep learning [4, 5, 6, 7, 8] has gained increasing attention and has been widely applied to various tasks in medical segmentation. The 3D encoder-decoder architecture [2, 9] with skip connections, enabling the effective fusion of low-level spatial details and high-level semantic features, has demonstrated strong performance in biomedical image segmentation. To further enhance the model's ability to capture critical anatomical structures, attention mechanisms [10] have been integrated into the encoder-decoder framework, improving robustness in complex and heterogeneous tumor regions. More recently, Transformer-based [11] architectures have been ex-

plored to capture long-range dependencies and global contextual information, further advancing the segmentation performance in multi-modality medical imaging. The effectiveness of these methods is largely attributed to the complementary information provided by multi-modality MRI data.

Approaches for incomplete multi-modality brain tumor segmentation, designed to address scenarios with limited modalities can be divided into modality completion [12, 13, 14] knowledge distillation [15, 16] and cross-modality learning [17, 18]. Learn cross-modality knowledge and share it with a latent space [13, 14] achieve good performance, but they lead to inconsistent performance when the missing modalities vary across samples. Recently, some knowledge distillation [15, 16] methods have been proposed to tackle this challenging problem that there is only one modality available at inference time. They aim to transfer knowledge from the teacher model trained using multi-modality images to the student model that is only trained with one modality. This strategy improves generalization and a regularization effect. However, knowledge distillation methods for limited modality scenarios often overlook critical modality-specific information, leading to lower accuracy. As a result, their generalization capability is limited when handling diverse patterns of missing modality.

In this paper, we propose a novel 3D MRI image segmentation architecture, EMSA-Net, to fully exploit the effective feature of multi-modality images for brain tumor segmentation. Inspired by knowledge distillation, we design a knowledge distillation framework that consists of a teacher model and a student model for a lightweight segmentation model. The teacher model is trained using all four modalities, while the student model is trained using 15 different combinations of modalities. This approach allows the student model to learn from the teacher model's knowledge, enabling effective segmentation even with limited modalities. Given the heterogeneity in lesion sizes and shapes, a novel Equivalent Multi-Scale Module (EMS) is needed to capture the relationship between different modalities and tumor regions. The EMS module consists of two components: feature extraction and modality attention. The feature extraction component employs a hierarchical cascade convolutional structure to extract modality-specific features at multiple scales, while the modality attention component dynamically generates attention maps weighted by the contributions of each modality during the down-sampling process. This design allows the network to focus on the most relevant features for accurate segmentation. To alleviate gradient disappearance and explosion issues, the Cross-Modality Attention Mechanism (CMA)

module employs residual learning. Inspired by attention, the CMA module incorporates both spatial and channel attention mechanisms to refine feature fusion within the skip connections. This mechanism enhances the network's ability to exploit complementary information across modalities, even when certain modality data is unavailable, yielding more discriminative features for accurate segmentation. We evaluate our method on BraTS2020 challenge datasets. The performance of the proposed model is superior to state-of-the-art methods.

Based on the above description, this paper focuses on the effect of different modalities on different areas of the brain tumor. The main contributions can be summarized as follows:

1. We propose a 3D MRI image segmentation method called EMSA-Net which integrates knowledge distillation to enhance performance under limited conditions. Specifically, the knowledge distillation framework is established between a teacher network and a student network, where the teacher network adopts a deeper architecture with a five-layer EMSA-Net to provide rich and informative supervision signals, while the student network utilizes a more lightweight four-layer design to ensure efficient inference.
2. The EMS module is specifically designed to capture the relationship between different modalities and tumor regions. Initially, hierarchical cascade convolutions extract modality-specific features at multiple scales. Subsequently, dynamically generated attention maps weighted by modality contributions are computed during the down-sampling process, guided by their diagnostic significance. Additionally, Group normalization is implemented to ensure stable training, particularly beneficial when dealing with small batch sizes.
3. The CMA mechanism is developed to refine feature fusion by incorporating both spatial and channel attention within the skip connections. This mechanism strengthens the network's ability to exploit complementary information across modalities, even when certain modality data is unavailable, yielding more discriminative features for accurate segmentation.
4. Experimental results demonstrate improvements in segmentation accuracy, highlighting the robustness and generalizability of our method under limited modality conditions. Comprehensive experiments using BraTS2020 challenge datasets evaluate the effectiveness of the proposed

method. It utilizes knowledge distillation, where a teacher model is trained with all four complete modalities and a student model is processed with 15 distinct input modality combinations to address limited modalities.

This paper is organized as follows: Section 2 provides a detailed overview of related work in medical image segmentation and brain tumor segmentation with limited modalities. Section 3 presents the proposed EMSA-Net architecture, including the EMS module, CMA module, and knowledge distillation framework. Furthermore, Section 4 describes the dataset and preprocessing, experiments details and evaluation metrics, and discusses the experimental results and ablation studies. Finally, Section 5 concludes the paper and outlines future research directions.

2. Realted Work

In this section, this paper provides a detailed introduction to the existing work on deep learning for medical image segmentation and brain tumor segmentation with limited modalities.

2.1. Deep learning in medical image segmentation with multi-modality

Deep learning [19, 20, 21, 22, 23, 24, 25] plays an important role in brain tumor segmentation. More recently, U-Net [26] introduced an encoder-decoder structure with skip connections to combine low-level spatial information from the encoder with high-level semantic information from the decoder, achieving remarkable success in biomedical image segmentation tasks. However, U-Net has poor robustness due to insufficient data in brain tumor segmentation. Attention U-Net [10] incorporated attention mechanisms to focus on the most relevant regions, improving the model's robustness in handling complex medical images. UNet3+ [9] emphasized multi-scale learning and hierarchical structures to integrate spatial details with a global context, achieving superior performance in complex and challenging medical image segmentation tasks. However, these models ignore the clinical scenarios where there may be missing modalities. Transformer-based [11, 21] architectures have been explored to capture long-range dependencies and global contextual information, further advancing the segmentation performance in multi-modality medical imaging. However, these methods often suffer from high computational complexity, limited generalization on small datasets, and inadequate modeling of local fine-grained features.

The attention mechanism [27, 28] has become widely used in medical image segmentation due to its ability to enhance model performance by focusing on the most relevant features in multi-modality MRI brain tumor datasets. One common form is the spatial attention mechanism [29], which assigns weights to spatial locations in an image to highlight critical regions, such as tumors, thereby improving segmentation accuracy. Additionally, SENet [30] introduced channel-wise attention by modeling inter-dependencies between channels. In this approach, global average pooling is used to compute channel-wise attention within the Squeeze-and-Excitation module. Furthermore, studies [31, 32] extended this idea by combining spatial and channel attention in a unified framework, improving the network's ability to focus on both critical spatial regions and informative feature channels. However, they failed to fully integrate spatial and channel attention, resulting in suboptimal feature fusion.

2.2. Brain tumor segmentation with limited modalities

MRI is widely used for brain tumor segmentation due to its ability to provide complementary information across different imaging modalities. However, in clinical scenarios, it is common for one or more modalities to be missing [33, 34, 35, 36, 37, 38] due to various constraints or patient-specific factors. For instance, the T1ce modality requires patients to ingest contrast agents (e.g., gadolinium-based agents), but some individuals are allergic to these agents. To address this challenge, researchers have explored several strategies, which can be broadly categorized into modality completion, knowledge distillation, and cross-modality learning.

Modality completion focuses on reconstructing missing modalities [12, 13, 39] which often requires training a generative model to generate missing modalities. Shen et al. [12] proposed a multi-domain image completion framework that handles random missing input data by exploiting modality relationships across multiple domains. Lee et al. [13] introduced CollaGAN, a collaborative GAN that imputes missing image data using all available modalities as mutual references. However, these methods typically require high-quality paired datasets for additional training and substantial computational resources.

Knowledge distillation [15, 16, 40, 41] has emerged as a powerful tool for addressing limited modalities in a lightweight manner. Hu et al. [16] distilled semantic and structural knowledge from a multi-modal teacher to a mono-modal student for segmentation. Chen et al. [15] leveraged privileged

multimodal information during training to enhance unimodal segmentation performance at test time. However, existing knowledge distillation methods focus solely on the modality available at inference time, which leads to reduced robustness in highly heterogeneous datasets.

Cross-modality learning [42, 43, 44] investigates inter-modal relationships to handle missing data during segmentation tasks. Recently, RFNet [17] adaptively fused incomplete modalities at a sub-region level into a shared latent space. However, its performance deteriorates sharply when only one modality is accessible. PPIL [45] introduced interaction mechanisms between pre-modality and post-modality representations, using contrastive learning and joint optimization to improve segmentation accuracy in the presence of missing MRI modalities. Nonetheless, accurately modeling inter-modality correlations remains challenging, particularly when the missing modality classes vary.

These strategies effectively provide valuable solutions for limited modality conditions, but they still face notable challenges, such as reliance on high-quality paired data, and lower accuracy.

3. Method

This section presents a detailed description of the proposed EMSA-Net architecture, with its schematic representation shown in Figure 2. Firstly, we delineate the hierarchical organization and information flow pattern of the network. Then, the novel EMS and CMA modules are analyzed, with emphasis placed on their core components. Finally, we demonstrate the knowledge distillation framework which effectively enhances segmentation performance in scenarios with limited imaging modalities.

3.1. Overview of EMSA-Net

Figure 2(a) illustrates the flowchart of the proposed EMSA-Net, a framework optimized for hierarchical feature extraction in medical imaging analysis. EMSA-Net adopts a U-Net-inspired encoder-decoder architecture with two key enhancements. In the encoding stage, a novel EMS module is introduced to perform multi-scale feature aggregation using modality-specific attention gates, which dynamically adjust contribution weights across available imaging modalities. This design enables robust feature representations, especially in scenarios involving incomplete data or missing critical imaging

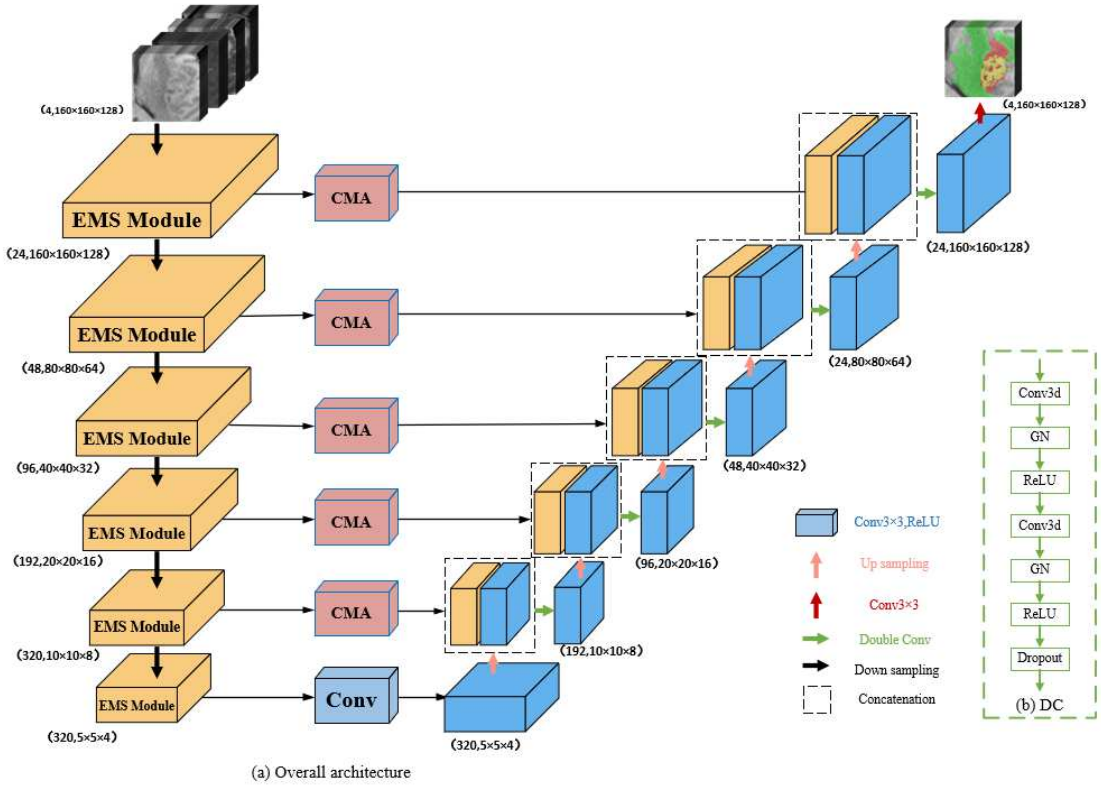


Figure 2: (a) Overall architecture of the proposed network in the knowledge distillation framework for brain tumor segmentation, consisting of multiple EMS modules followed by CMA blocks. The network performs down-sampling and up-sampling operations using convolutional layers and transposed convolutions to generate the final segmentation output. (b) Double convolution block.

sequences. Furthermore, an innovative CMA module is strategically integrated into the skip connections, bridging the feature hierarchies between the encoder and decoder. The CMA module further refines feature fusion by leveraging spatial and channel attention mechanisms to highlight the most relevant features from each modality. This maintains structural consistency and enables the model to dynamically adapt to limited modality situations during both the down-sampling and up-sampling stages. In the decoding stage, a double convolution block is employed, consisting of two consecutive 3D convolutional layers, each followed by group normalization (8 groups) and ReLU activation, as illustrated in Figure 2(b).

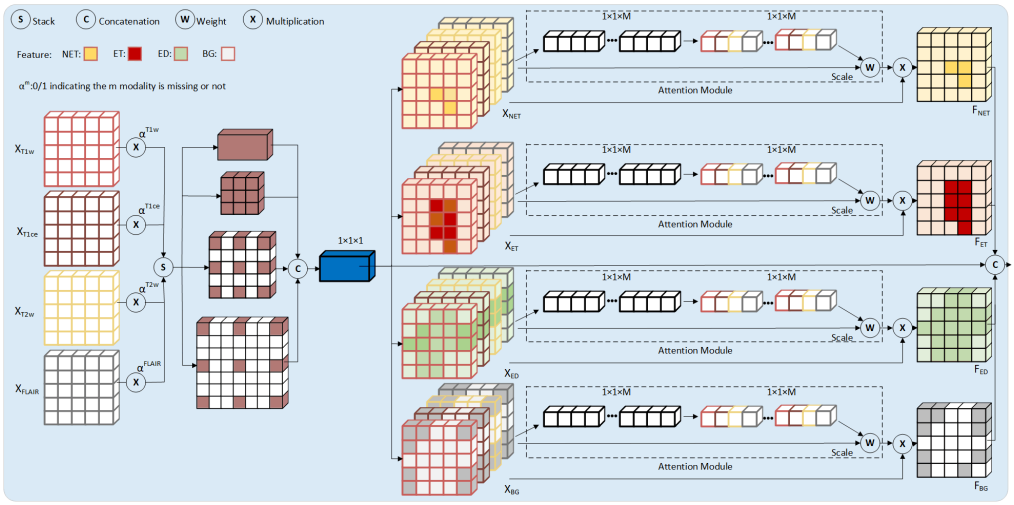


Figure 3: The EMS module for down-sampling, including feature extraction and modality attention for multi-modality is present. Feature extraction for fully extracting modality-specific features, and modality-attention for adaptively weighing the contribution of modality to brain tumor regions.

3.2. Equivalent Multi-Scale Module

To comprehensively exploit complementary information from multi-modality images, this paper proposes an EMS module that dynamically allocates varying attention weights to different imaging modalities across anatomical regions. Specifically, the design motivation stems from two critical observations: First, traditional convolutional operations with fixed receptive fields struggle to capture fine-grained details of small tumors and the overall contours of large tumors simultaneously in brain MRI, given the substantial heterogeneity in lesion sizes and shapes. Second, native integration of multi-scale convolution architectures significantly increases parameter complexity and computational cost, rendering them clinically impractical for real-time diagnostic applications. To address these limitations, the EMS module employs a computationally efficient architecture comprised of two synergistic components: feature extraction and modality attention.

Feature Extraction. A hierarchical feature extractor with adaptive receptive fields is proposed to maximize the utilization of information from available imaging modalities, thereby achieving precise tumor region delineation. Given the features $X_m \in \mathbb{R}^{C\times H\times W\times D}$ from each imaging modality $m \in M$, this paper implements enhanced multi-scale convolution to syn-

ergistically extract and integrate complementary features across modalities through hierarchical cascade convolutions. As illustrated in Figure 3, we define a gating parameter α^m set to either 0 or 1, where 0 indicates the absence of the m-th modality, and 1 indicates its presence. The feature extraction procedure is defined as:

$$X_r{}' = ⓒ \sum\nolimits_{m \in M} f^{fe} \cdot \alpha^m \cdot x^m, \tag{1}$$

where ⓒ represents the concatenation operation. M denotes the modality set, including T1w, T1ce, T2w, and FLAIR, and f^{fe} is the feature extraction. To reduce dimensionality and enhance computational efficiency, $X_r{}'$ is processed to X_r through a bottleneck convolutional layer, expressed as

$$X_r = Conv_{1\times1\times1} X_r{}', \tag{2}$$

where r represents brain tumor regions and $r \in R$, including NET, ET, ED, and BG. X_r highlights the distinct contributions of each modality to different brain tumor regions, capturing the unique information provided by each imaging technique. This enables the model to more effectively integrate multi-modality data and leverage the strengths of each modality in segmenting various tumor structures.

Modality Attention. After feature extraction, the feature maps from various modalities are classified into four brain tumor regions: NET, ET, ED, and BG. Subsequently, a modality-aware attention mechanism that prioritizes clinically salient features helps the network better understand the unique contributions of each modality. Through this mechanism, the network dynamically assigns higher attention weights to clinically significant features, which are then multiplied with the corresponding feature maps to emphasize critical information. The modality attention process can be expressed as

$$F_r = FM \cdot Attention(X_m), \tag{3}$$

where F_r is the fused feature map. The fused feature map is supervised by a weighted cross entropy loss L_{WCE} and a Dice loss L_{Dice}, expressed as

$$L_{F_r} = \sum\nolimits_{r \in R} (1-\beta) \cdot L_{WCE}(p_r, y) + \beta \cdot L_{Dice}(p_r, y), \tag{4}$$

where p_r is the prediction of brain tumor region and y represent the ground truth label. Weight for the region can be represented as $w_r = \frac{N - \sum_{r \in R} p_r}{\sum_{r \in R} p_r}$, and

L_{WCE} is calculated by

$$L_{WCE}(p_r, y) = \sum\nolimits_{r \in R} w_r \cdot y \cdot lg(p_r). \tag{5}$$

L_{Dice} is expressed as

$$L_{Dice}(p_r, y) = 1 - \sum\nolimits_{r \in R} \frac{2|p_r \cap y|}{(|p_r| + |y|) \cdot R_{in}}, \tag{6}$$

where R_{in} is the number of input regions.

To address the inherent limitations of batch normalization under limited batch size conditions, the EMS module adopts Group Normalization as a memory-efficient alternative. Additionally, ReLU activation is employed to stabilize training and improve overall performance. By incorporating these innovations, the EMS effectively addresses domain-specific challenges related to partial modality availability, offering a robust and efficient solution for medical image segmentation workflows.

3.3. Cross-Modality Attention Mechanism

To further enhance multi-modality feature fusion and mitigate the impact of missing modalities in medical image segmentation, we propose the CMA module, which is shown in Figure 4. This module establishes bidirectional cross-talk between encoder and decoder pathways through skip connections and leverages a mixed attention mechanism to refine feature representation.

The CMA module utilizes a hierarchical attention mechanism to prioritize the most relevant features from each modality. The mechanism first processes channel-wise feature information, and then refines spatial features enhancement, establishing complementary local-global contextual integration. This design enhances the robustness of feature representation under limited data conditions.

As illustrated in Figure 4, the CMA module comprises three principal components: 1) Cross-fusion Block (CFB) for multi-scale feature integration through skip-connections and up-sampling operations. 2) Channel-attention block (CAB) utilizes dual-pooling aggregation (global average pooling and max pooling) with adaptive weights for each channel via a lightweight fully connected network. and 3) Spatial-attention block (SAB) generates position-sensitive attention maps through channel-wise pooling and convolutional processing.

Algorithm 1: Channel-Modality Attention (CMA) Module

Input: Input feature map x; Iteration number N, learning rate η

Output: Output feature map y after CMA module

for $i = 1$ **to** N **do**

    // Step 1: Channel Fusion Block (CFB)

    $conv_1 = \text{Conv3D}(x)$;

    $conv_2 = \text{Conv3D}(x)$;

    $gn_1 = \text{GroupNorm}(conv_1)$;

    $gn_2 = \text{GroupNorm}(conv_2)$;

    $cfb_output = \text{ReLU}(gn_1 + gn_2)$;

    // Step 2: Channel Attention Block (CAB)

    $avg_pool = \text{AdaptiveAvgPool}(cfb_output)$;

    $max_pool = \text{AdaptiveMaxPool}(cfb_output)$;

    $avg_fc = FC(\text{ReLU}(FC(avg_pool)))$;

    $max_fc = FC(\text{ReLU}(FC(max_pool)))$;

    $channel_att = \sigma(avg_fc + max_fc)$; // Channel attention weights

    $cab_output = cfb_output \cdot channel_att$; // Apply channel attention

    // Step 3: Spatial Attention Block (SAB)

    $avg_channel = \text{Mean}(cab_output, \text{dim=channel})$;

    $max_channel = \text{Max}(cab_output, \text{dim=channel})$;

    $spatial_concat = \text{Concat}(avg_channel, max_channel, \text{dim=channel})$;

    $spatial_conv = \text{Conv3D}(spatial_concat)$;

    $spatial_att = \sigma(spatial_conv)$; // Spatial attention weights

    $y = cfb_output \cdot cab_output \cdot spatial_att$; // Final attention output

    $x = y$; // Update input for next iteration

return y

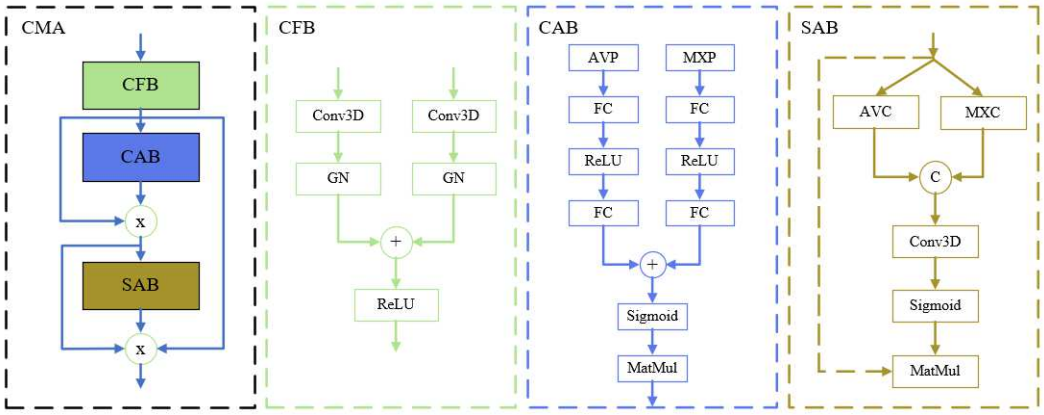


Figure 4: The Cross-Modality Attention Mechanism (CMA) is strategically integrated into the skip connections, bridging the feature hierarchies between the encoder and decoder. It includes three main blocks: the Cross-feature block (CFB), the Channel-attention block (CAB), and the Spatial-attention block (SAB).

The hierarchical attention outputs are then fused with encoder-decoder skip connections using learnable fusion coefficients, followed by convolutional processing to generate unified cross-modality representations. This dual-attention mechanism with residual learning not only improves the network's ability to integrate features from different modalities but also enhances its robustness to limited modalities, demonstrating superior performance on multi-modality MRI datasets. In addition, residual structures help alleviate gradient disappearance and explosion issues, ensuring stable gradient flow during backpropagation. Algorithm 1 is summarized to illustrate the training process of the CMA module.

3.4. Knowledge Distillation

In this study, a standard knowledge distillation framework is employed to enhance the performance of an incomplete modality segmentation model under limited modality scenarios. The general architecture is depicted in Figure 5. The teacher model, trained with complete modalities, guides the student model, which is trained on a subset of the teacher's inputs. The teacher model is pre-trained using all available modalities, effectively capturing rich and comprehensive multi-modal information.

During the training phase, the teacher model provides soft labels to guide the student model. The knowledge distillation loss is formulated using

Algorithm 2: Knowledge Distillation with Feature and Soft Target Loss

Input: Teacher model parameters θ_t, Student model parameters θ_s; Multi-modality input x^* (for teacher), Incomplete modality input x (for student); Ground truth y, Temperature T, weights α, β; Iteration number N, learning rate η

Output: Trained student model parameters θ_s

Initialize θ_s and θ_t randomly;

for $i = 1$ **to** N **do**

- $z^t = \text{TeacherEncoder}(x^*)$;
- $o^t = \text{TeacherDecoder}(z^t)$; `// Pre-train teacher model`
- $p^t = \text{Softmax}(o^t)$; `// Softmax with` $T = 1$
- $\mathcal{L}_{F_r} = (1-\beta)\,\mathcal{L}_{WCE}(p^t, y) + \beta \mathcal{L}_{Dice}(p^t, y)$; `// Segmentation Loss`
- $\theta_t \leftarrow \theta_t - \eta \nabla_{\theta_t} \mathcal{L}_{seg}$; `// Update teacher model`

for $j = 1$ **to** N **do**

- $z^s = \text{StudentEncoder}(x)$;
- $o^s = \text{StudentDecoder}(z^s)$;
- $p^s = \text{Softmax}(o^s)$; `// Student prediction with` $T = 1$
- $p^s_T = \text{Softmax}(o^s/T)$; `// Student output with` T
- $z^t = \text{TeacherEncoder}(x^*)$;
- $o^t = \text{TeacherDecoder}(z^t)$;
- $p^t_T = \text{Softmax}(o^t/T)$; `// Teacher soft target`
- $\mathcal{L}_{seg} = \mathcal{L}_{WCE}(p^t, y) + \mathcal{L}_{Dice}(p^t, y)$;
- $\mathcal{L}_{KD} = \text{KL}(p^s_T \parallel p^t_T)$; `// KD loss`
- $\mathcal{L}_{FL} = \text{FeatureLoss}(z^s, z^t)$; `// Feature loss`
- $\mathcal{L}_{total} = \mathcal{L}_{F_r} + \mathcal{L}_{seg} + \mathcal{L}_{KD} + \mathcal{L}_{FL}$;
- $\theta_s \leftarrow \theta_s - \eta \nabla_{\theta_s} \mathcal{L}_{total}$; `// Update student model`

return θ_s

Kullback-Leibler (KL) divergence:

$$L_{KD}(P_{teacher}||P_{student}) = \sum\nolimits_{y} P_{teacher}(y|x) \log(\frac{P_{teacher}(y|x)}{P_{student}(y|x)}), \tag{7}$$

which measures the divergence between the student's outputs and the teacher's outputs in both label space and feature space. The student model is trained to minimize a hybrid loss function that combines a supervised loss—measuring the accuracy of segmentation predictions on intermediate feature representations, written as L_{FL}, where uses mean squared error.

Student loss is calculated as

$$L_{seg}(P_r, y) = \sum\nolimits_{i=1}^{N} \sum\nolimits_{m \in M} L_{WCE}(P_r, y) + L_{Dice}(p_r, y), \tag{8}$$

where N is the number of training data.

Total loss is expressed as

$$L_{total} = L_{F_r} + L_{seg} + L_{KD} + L_{FL}. \tag{9}$$

The distillation loss enables effective cross-knowledge transfer from the teacher model to the student model, allowing the latter to replicate the teacher's predicative patterns despite being restricted to partial input modalities. This approach not only bridges the modality gap but also enhances the student model's generalization capability, making it suitable for clinical applications where data completeness is often a challenge. Algorithm 2 is summarized to illustrate the training process of the knowledge distillation framework.

4. Experiments and Results

4.1. Dataset and Preprocessing

This paper evaluates the proposed method on the BraTS2020 challenge dataset, which consists of 369 brain tumor cases with manually annotated labels. Each case contains four MRI modalities, including T1w, T2w, T1ce, and FLAIR. The manually annotated labels are divided into four categories: background (BG), necrotic and non-enhancing core (NCR/NET), peritumoral edema (ED), and enhancing tumor (ET). For evaluation, this paper follows the task setting in the challenge to merge the tumor classes into the whole tumor (WT) region including all tumor classes, tumor core (TC) region consisting of ET, NET, and enhancing tumor (ET) region.

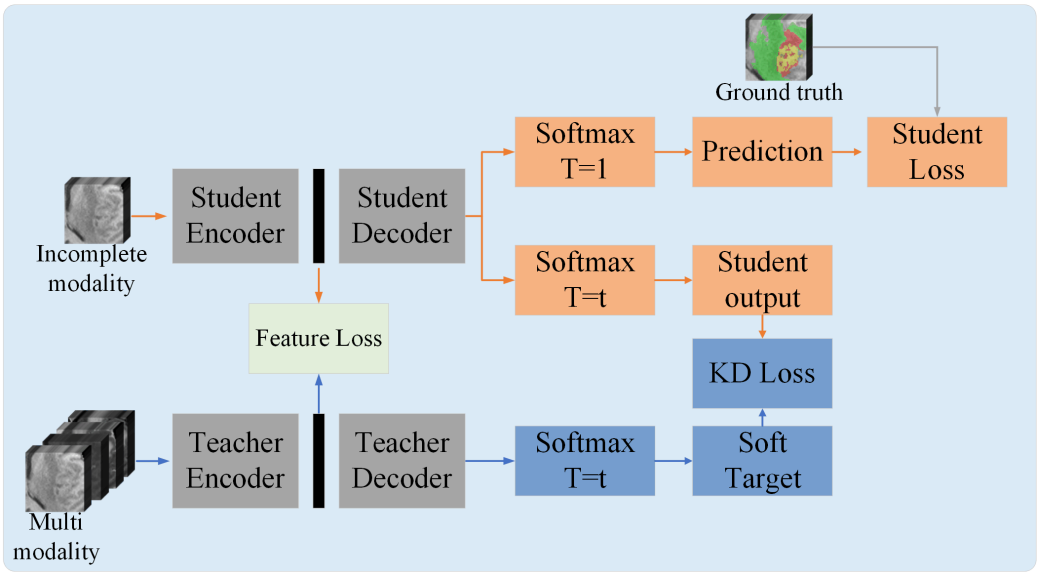


Figure 5: Overview of the knowledge distillation framework for brain tumor segmentation.

For pre-processing, each volume is normalized to have a mean of 0 and a standard deviation of 1 within the foreground region. This paper randomly crops each volume to feed the network due to limited GPU memory. Afterward, the BraTS2020 dataset is randomly split into 295 and 74 subjects for training and testing, respectively.

4.2. Experimental Details and Evaluation Metrics

All experiments in the article are conducted on a server with an Ubuntu 22.04 workstation equipped with an Intel(R) Core(TM) i9-12900KF CPU @5.20 GHz, 32 GB of RAM and a single NVIDIA GeForce RTX3090 GPU with 24 GB of RAM. The mainstream deep learning platform PyTorch is used to build our deep learning model, version 1.13. During model training, we employ backpropagation with an optimizer to iteratively update network parameters, aiming to minimize the loss function and achieve model optimization. Specifically, our proposed EMSA-Net is trained using the Adam optimizer, which dynamically adjusts each model parameter using first-order and second-order moment estimates of gradients. The training configuration consisted of an initial learning rate of 1e-3 and a group number of 8. The training was configured with a maximum of 500 epochs, and an early stopping strategy was employed with patience of 30 epochs, meaning that

training would stop early if no improvement was observed on the validation set for 30 consecutive epochs.

This paper first implements a high-performing teacher model using a designed encoder-decoder architecture. Subsequently, the teacher model guides a student model, which uses a subset of the teacher's input data to effectively learn feature representations. During inference, the student model achieves accurate predictions without requiring the privileged multi-modality data available to the teacher.

To comprehensively evaluate the performance of the proposed method, Dice Score (DSC) is employed, measuring the voxel-wise accuracy. DSC measures overlap between prediction and ground truth, and higher DSC indicates better performance.

In the ablation study for the teacher model, the DSC, Jaccard coefficient (JC), 95% Hausdorff distance (HD95), and average surface distance (ASD) are used to evaluate the effectiveness of key components in brain tumor segmentation. DSC measures the voxel-wise accuracy; JC offers a stricter evaluation than Dice by penalizing segmentation errors more; Hd95 measures the extreme boundary error by considering the 95th percentile of the Hausdorff distance; and ASD measures the average boundary error between the predicted and ground truth surfaces. For the student network, DSC, sensitivity, and HD95 are used to evaluate the effectiveness of the components. Sensitivity measures the proportion of true positive predictions among all actual positive cases, providing insight into the model's ability to identify tumor regions accurately.

4.3. Comparisons with the State-of-the-art

In this section, the performance of the proposed EMSA-Net for brain tumor segmentation is evaluated. We benchmark our method against five state-of-the-art approaches including (1) PPIL [45] (2) RFNet [17] (3)RobustSeg [42] (4)MFI [46], and (5) M2FTrans [14]. Quantitative comparison results under fifteen multi-modality combinations on the BraTS2020 dataset are summarized in Table 1. Based on the experimental results presented above, observations are as follows.

1. Among the comparison approaches, M2FTrans [14] achieves the best overall segmentation performance for WT, TC, and ET, outperforming PPIL [45] under 28 of 45 modality settings. 45 represents the results of the DSC, in which 15 multi-modality combinations multiply 3 types (WT, TC, and ET) of brain tumors.

Table 1: Quantitative comparison results measured in DSC (%) on BraTS2020.

Type	FLAIR	T1w	T1ce	T2w	PPIL [45]	RFNet [17]	RobustSeg [42]	MFI [46]	M2FTrans [14]	Ours	Average
WT				√	87.6	86.0	82.2	85.3	87.2	**88.4**	87.9
			√		80.4	76.7	71.3	73.1	78.8	**80.6**	86.9
		√			81.1	77.1	71.4	69.4	79.1	**81.5**	84.1
	√				88.6	87.3	82.8	85.1	88.7	**89.5**	85.5
			√	√	88.2	87.7	85.9	88.0	88.6	**89.8**	85.5
		√	√		82.7	81.1	76.8	77.4	82.4	**83.2**	89.4
	√	√			89.6	89.7	88.1	89.2	**90.3**	89.4	87.8
	√		√		88.1	87.7	85.5	86.9	88.3	**88.9**	88.9
	√			√	90.4	89.8	88.0	89.5	90.5	**90.6**	87.8
	√		√		90.0	89.8	87.3	89.2	90.3	**91.5**	88.7
	√	√	√		90.3	90.6	88.8	90.0	91.0	**91.8**	88.7
	√	√		√	90.6	90.6	89.2	90.2	90.9	**91.2**	88.7
	√	√	√		90.8	90.6	88.6	90.5	91.1	**92.4**	88.7
	√		√	√	88.4	88.3	86.6	87.8	89.0	**89.5**	88.7
	√	√	√	√	91.0	91.1	89.4	90.8	91.3	**92.7**	**88.7**
TC				√	73.1	71.0	61.8	66.6	72.3	**75.3**	79.8
			√		84.4	81.5	76.6	78.5	81.8	**88.0**	78.2
		√			68.2	66.0	54.3	54.7	66.7	**70.5**	73.4
	√				72.0	69.1	60.7	65.0	72.2	**75.1**	76.0
			√	√	86.1	83.4	82.4	84.3	84.6	**90.0**	79.2
		√	√		84.8	83.4	80.2	81.6	83.7	**88.9**	85.4
	√	√			72.8	73.0	68.1	71.5	74.4	**75.7**	85.4
	√		√		74.2	73.1	66.4	69.5	73.6	**76.6**	85.4
	√			√	75.4	74.1	68.2	71.2	75.4	**75.5**	85.4
	√		√		85.7	84.6	81.8	83.9	85.5	**89.1**	85.4
	√	√	√		85.7	85.0	82.7	85.2	85.8	**88.8**	85.4
	√	√		√	75.6	75.1	70.4	72.5	76.1	**77.6**	85.4
	√	√	√		86.3	84.9	81.8	84.8	85.2	**90.3**	85.4
	√		√	√	86.0	83.5	82.8	84.8	84.9	**89.0**	85.4
	√	√	√	√	86.3	85.2	82.8	85.5	85.4	**90.4**	**82.7**
ET				√	50.3	46.2	36.4	43.5	51.5	**53.1**	65.8
			√		81.8	74.8	67.9	74.0	82.5	**83.2**	61.4
		√			40.3	37.3	28.9	29.6	40.8	**44.0**	55.4
	√				44.6	38.1	34.6	39.4	43.3	**54.1**	62.2
			√	√	81.1	75.9	71.4	80.1	82.3	**82.8**	66.7
		√	√		81.8	78.0	70.1	78.3	**83.8**	81.3	**84.3**
	√	√			44.4	40.9	39.6	42.8	47.0	**54.1**	**84.3**
	√		√		49.7	45.6	39.9	47.1	49.9	**56.2**	**84.3**
	√			√	53.1	49.3	42.1	46.7	53.8	**54.7**	**84.3**
	√		√		82.1	76.6	70.7	80.1	**83.0**	82.3	**84.3**
	√	√	√		82.1	76.8	71.7	81.4	**84.1**	81.2	**84.3**
	√	√		√	52.4	49.9	43.9	47.0	53.3	**60.4**	**84.3**
	√	√	√		81.2	77.1	71.1	80.4	81.2	**81.7**	**84.3**
	√		√	√	81.1	77.0	71.8	80.4	82.3	**83.0**	**84.3**
	√	√	√	√	81.3	78.0	71.5	81.9	82.1	**84.3**	**84.3**

2. Comparatively, the proposed model consistently outperforms all five methods under 41 out of 45 modality settings for WT, TC, and ET and achieves an average increase of 0.8%, 2.9%, and 2.1% in average DSC respectively.
3. With single modality, the proposed model achieves significant improvements over other methods. For instance, when using only T1w, the performance of the proposed model increases by 0.4%, 2.3%, and 3.2% for WT, TC, and ET respectively. This indicates that the proposed method effectively leverages information from a single modality to enhance segmentation accuracy.
4. Additionally, the proposed model outperforms other methods in combining limited modalities. For instance, in the combination of T1ce and T2w, the performance of the proposed model increases by 1.2%, 3.9%, and 0.5% for WT, TC, and ET respectively. The data analysis confirms that the performance of the proposed method improves with incomplete modalities.
5. Furthermore, when all four modalities are available, the proposed model achieves DSC scores of 92.7%, 90.4%, and 84.3% for WT, TC, and ET respectively, surpassing the best-performing M2FTrans [14] by 1.4%, 4.1%, and 2.1%. This demonstrates that the proposed method not only excels in limited modality scenarios but also maintains superior performance when complete modality data is accessible.

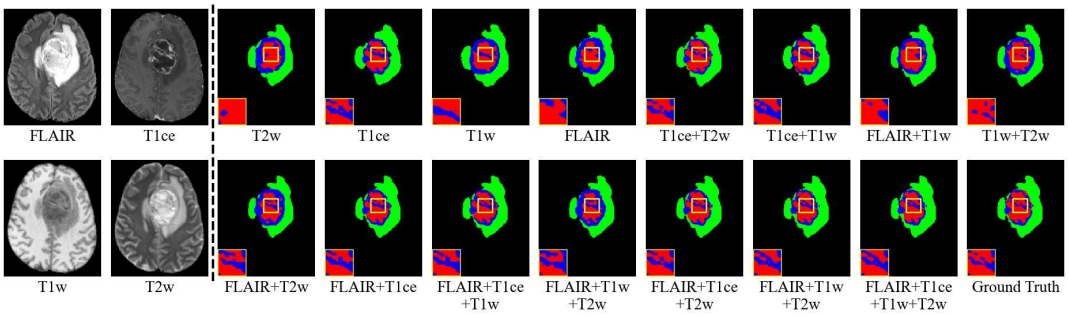


Figure 6: Qualitative segmentation results of our architecture on BraTS2020 under all fifteen multi-modality settings.

This paper further visualizes the segmentation results of our network across all fifteen multi-modality combinations, demonstrating that the proposed method effectively segments brain tumors in various limited scenarios, as shown in Figure 6 (red indicates ET, blue indicates TC, and green indicates WT). The visualization highlights the segmentation of TC, particularly through the enlarged patches in the lower-left corner of each prediction. The region is often the most challenging to segment accurately due to its complex structure and clinical importance. When only one modality is available, T1ce proves to be the most informative modality for the segmentation of TC. The combination of T1ce and T2w further enhances the segmentation accuracy, producing results visually more consistent with the ground truth. As more modalities are added, segmentation results become comparable with complete modalities. These results validate the effectiveness of the proposed model in research involving limited modality.

4.4. Ablation study

To get a better understanding of the effectiveness of key components in the proposed method, this paper conducts two types of ablation studies. First, an ablation is conducted for the teacher network, representing the overall contribution of each component to brain tumor segmentation. Second, ablations are conducted for the student network under limited modality, demonstrating the effectiveness of each component in the case of limited modalities.

In detail, four configurations are implemented to validate the contribution of different key components: 1) a baseline network is trained; 2) an EMS module is added to the baseline network; 3) a CMA mechanism alone is

incorporated into the baseline network; 4) both the EMS module and CMA mechanism are integrated to complete the framework.

Table 2: Component-wise ablation study for the teacher network on BraTS2020 under four complete modalities.

Modules	Metrics			
	Dice(%)↑	JC(%)↑	HD95(mm)↓	ASD(mm)↓
Baseline	91.5	84.8	6.956	2.201
Baseline+EMS	91.4	84.8	4.792	1.431
Baseline+CMA	**92.3**	**86.0**	4.503	1.192
Baseline+EMS+CMA	92.1	85.6	**4.320**	**0.953**

Table 2 represents the ablation for the teacher network with four complete modalities, and the following observations can be made.

1. The baseline model achieves a Dice coefficient of 91.5% and a JC of 84.8%, but its boundary-related metrics, including an HD95 of 6.956 and an ASD of 2.201, reveal significant limitations in boundary precision. When the EMS module is incorporated, there is a notable improvement in boundary metrics, with HD95 reduced to 4.792 and ASD reduced to 1.431, indicating the module's strength in refining segmentation boundaries.
2. Similarly, the CMA module enhances segmentation accuracy and boundary quality, improving the Dice coefficient to 92.3% and JC to 86.0% while reducing HD95 to 4.503 and ASD to 1.192. This highlights the CMA module's ability to improve overall segmentation consistency and precision.
3. The integration of both modules achieves the most balanced and optimal performance, with HD95 further reduced to 4.320 and ASD to 0.953, reflecting a 37.9% and 56.7% reduction from the baseline, respectively. While there is a marginal decrease in Dice and JC compared to the CMA module alone, the boundary precision shows a synergistic improvement, underscoring the complementary nature of these two modules. These results validate the research significance of the EMS and CMA modules in improving segmentation accuracy and boundary delineation, particularly in scenarios where precise boundary definition is critical.

Based on the analysis of the above experiments, it can be concluded that all the components of the proposed network play an indispensable role in brain tumor segmentation.

Table 3: Component-wise ablation study for Dice Score (DSC %) on BraTS2020.

Modality	Region	Baseline	Baseline+EMS	Baseline+CMA	Baseline+EMS+CMA
T2w	ET	46.8	53.0	**53.1**	**53.1**
	TC	68.6	73.9	74.5	**75.3**
	WT	81.7	87.5	88.2	**88.4**
	Average	65.7	71.5	71.9	**72.3**
T1ce+T2w	ET	80.6	80.5	82.0	**82.8**
	TC	84.3	88.2	89.7	**90.0**
	WT	84.9	89.2	89.6	**89.8**
	Average	83.3	86.0	87.1	**87.5**
FLAIR+T1ce+T2w	ET	**82.0**	80.1	80.4	81.7
	TC	85.9	88.3	89.8	**90.3**
	WT	90.3	92.2	91.9	**92.4**
	Average	86.1	86.9	87.4	**88.1**
FLAIR+T1ce+T2w+T1w	ET	81.8	81.4	82.1	**84.3**
	TC	83.9	89.5	90.1	**90.4**
	WT	90.4	92.5	92.3	**92.7**
	Average	85.4	87.8	88.1	**89.1**

Table 4: Complexity of the proposed method on BraTS2020 under limited modality.

Modules	Params	GFLOPs
Baseline	20.92	307.15
Baseline+EMS	+4.10	-14.65
Baseline+CMA	-0.02	+1.56
Baseline+EMS+CMA	+12.31	+25.95

Table 3 represents the ablation study for the student network under limited modality, and Table 4 shows the complexity changes of the components. The results are presented for four different scenarios: T2w, T1ce+T2w, FLAIR+T1ce+T2w, and FLAIR+T1ce+T2w+T1w. Based on the experimental results presented above, we have the following observations.

1. The baseline model struggles when dealing with limited modalities, particularly in the T2w-only scenario, where the average DSC is only 65.7%. This highlights the challenge of segmenting brain tumors with limited modality information.
2. The EMS module significantly improves segmentation in limited modality scenarios by better capturing multi-scale tumor features. In the

T2w-only setting, the DSC improves by 5.8%. Particularly, for T1ce+T2w, performance increases from 83.3% to 86.0%. Despite adding 4.10M parameters, EMS notably reduces the computational cost by 14.65 GFLOPs, making feature extraction more efficient.

3. In the T2w-only setting, the DSC reaches 71.9%, demonstrating CMA's ability to compensate for limited modalities. In other combinations, the performance also has a lot of improvement, highlighting the effectiveness of the CMA module in modality fusion. Meanwhile, CMA maintains a lightweight design (-0.02M parameters) and increases only a minor computational overhead (+1.56 GFLOPs), making it an efficient enhancement.
4. When both EMS and CMA are integrated, segmentation accuracy is maximized across all limited scenarios. Specifically, the highest DSC of 72.3%, 87.5%, 88.1% and 89.1% are achieved in the T2w, T1ce+T2w, FLAIR+ T1ce+T2w, and FLAIR+ T1ce+T2w+T1w scenarios, respectively. Additionally, the combined approach only adds 12.31M parameters and increases GFLOPs by 25.95, demonstrating a trade-off between accuracy and computational efficiency.

In conclusion, the proposed method effectively addresses limited modality scenarios by leveraging EMS for enhanced feature extraction and CMA for improved cross-modality feature fusion, ensuring robust segmentation despite limited inputs.

Sensitivity is a crucial metric for evaluating the performance of the student network, particularly in the context of limited modality use. It measures the model's ability to accurately identify true positive cases, providing insight into its effectiveness in detecting tumor regions. The results of the sensitivity are presented in Table 5, which highlights the performance of the student network across different modality combinations.

1. The baseline model, without any additional modules, demonstrates a steady improvement in sensitivity as more modalities are added. However, its performance is relatively limited in the single-modality setting, with an average sensitivity of only 0.655 under T2w input.
2. Introducing the EMS module leads to noticeable improvements across most modality settings, particularly in single-modality and dual-modality inputs. For instance, under T2w-only input, the average sensitivity increases from 0.655 to 0.738, indicating that EMS helps compensate for

Table 5: Component-wise ablation study for Sensitivity on BraTS2020.

Modality	Region	Baseline	Baseline+EMS	Baseline+CMA	Baseline+EMS+CMA
T2w	ET	0.491	0.585	0.626	**0.628**
	TC	0.624	0.747	**0.755**	0.717
	WT	0.850	0.881	0.849	**0.890**
	Average	0.655	0.738	0.743	**0.745**
T1ce+T2w	ET	0.827	0.851	0.856	**0.890**
	TC	0.844	0.864	0.848	**0.885**
	WT	0.875	0.875	**0.890**	0.885
	Average	0.849	0.863	0.865	**0.886**
FLAIR+T1ce+T2w	ET	0.834	0.852	0.848	**0.875**
	TC	0.839	0.846	0.854	**0.874**
	WT	**0.921**	0.903	0.917	0.920
	Average	0.865	0.867	0.873	**0.890**
FLAIR+T1ce+T2w+T1w	ET	0.867	0.866	0.879	**0.880**
	TC	0.859	0.852	0.868	**0.886**
	WT	**0.910**	0.899	0.908	0.901
	Average	0.879	0.872	0.885	**0.889**

the lack of modality diversity by reinforcing hierarchical feature learning.

3. The CMA module also consistently improves sensitivity compared to the baseline, especially in the ET and TC subregions. It demonstrates strong generalization across all modality configurations, with the average sensitivity reaching 0.885 in the full-modality scenario — outperforming both the baseline.
4. The combined use of EMS and CMA modules yields the best overall performance in terms of sensitivity across all modality settings. The improvements are especially prominent in ET and TC regions, where precise segmentation is often more challenging. This suggests that EMS and CMA are complementary: EMS strengthens feature representation at multiple scales, while CMA facilitates effective cross-modal information fusion, together resulting in more accurate and sensitive tumor segmentation.

HD95 measures the maximum distance between the predicted and ground truth segmentation boundaries, providing insight into the model's ability to accurately delineate tumor regions. The results of the HD95 are presented in Table 6. Both the EMS and CMA modules significantly reduce HD95 values across most modality settings, indicating better boundary accuracy. Notably, CMA demonstrates strong improvements in low-modality settings, such as T2w, where it reduces the average HD95 from 9.369 to 5.271. When combined, EMS and CMA achieve the lowest HD95 scores in high-modality

Table 6: Component-wise ablation study for HD95 (mm) on BraTS2020.

Modality	Region	Baseline	Baseline+EMS	Baseline+CMA	Baseline+EMS+CMA
T2w	ET	10.846	7.489	8.618	**7.418**
	TC	12.105	6.642	5.901	**5.866**
	WT	8.322	3.080	**2.051**	2.372
	Average	9.369	5.503	5.547	**5.271**
T1ce+T2w	ET	**3.331**	3.838	4.363	3.953
	TC	3.470	4.292	**2.789**	2.909
	WT	2.723	2.365	**2.177**	2.282
	Average	2.976	3.544	3.130	**3.067**
FLAIR+T1ce+T2w	ET	5.125	5.029	**3.799**	3.863
	TC	4.576	3.617	2.505	**2.258**
	WT	2.415	2.524	**1.544**	1.669
	Average	4.036	3.742	2.623	**2.595**
FLAIR+T1ce+T2w+T1w	ET	3.863	3.973	3.444	**2.945**
	TC	4.155	2.748	2.451	**1.919**
	WT	2.301	**1.665**	1.777	1.703
	Average	3.059	2.768	2.553	**2.205**

scenarios, with an average of 2.205 under full modality input, highlighting their complementary roles in enhancing precise tumor delineation.

Figure 7 visualizes the effect of different components on different combinations of image modalities. The EMS module and the CMA module also demonstrate their effectiveness in improving segmentation performance. The combination of EMS and CMA modules further enhances segmentation accuracy, as evidenced by the improved visual consistency with ground truth across all modality combinations. This highlights the complementary nature of these two modules in enhancing segmentation performance.

4.5. Model efficiency

Table 7 is the comparison of the model complexity and Figure 8 visually demonstrates that the proposed model achieves a superior trade-off between segmentation accuracy and computational efficiency. Compared with state-of-the-art methods, our proposed approach exhibits higher GFLOPs (333.10GFLOPs) while maintaining a more compact parameter size (33.53M). Importantly, the proposed method achieves Dice scores of 88.7% for WT, 82.7% for TC, and 68.8% for ET, outperforming other methods by 0.8%, 2.9%, and 2.1% respectively. Notably, the proposed approach holds significant potential for enhancement through targeted lightweight design strategies.

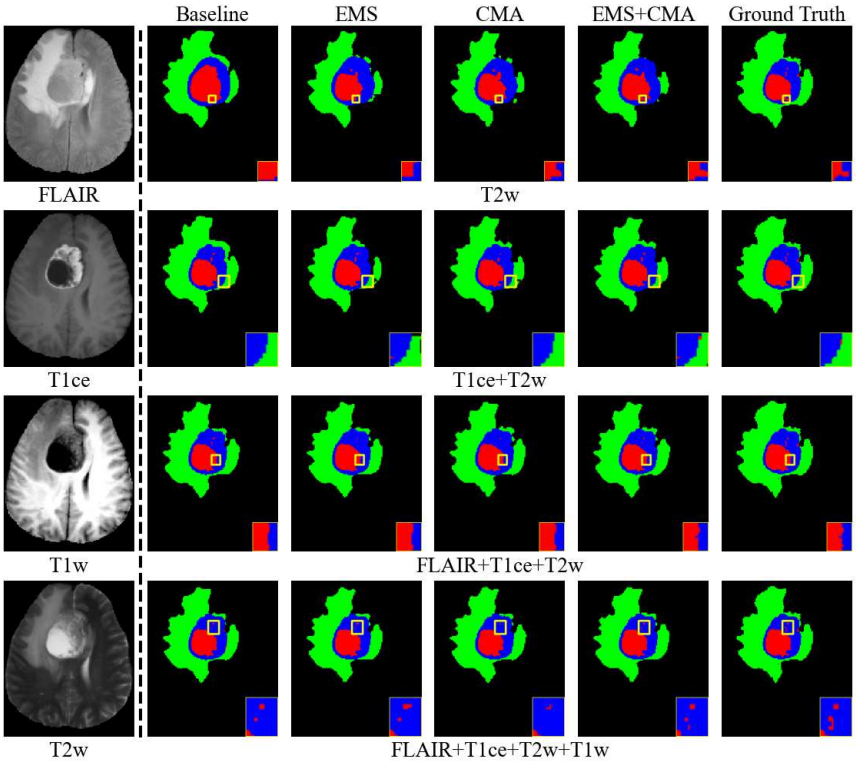


Figure 7: Ablation to different combinations of image modalities. Left: four image modalities. Right: segmentation masks predicted from different combinations of image modalities and the corresponding ground truth.

5. Conclusion

This paper proposes the EMSA-Net, a specialized architecture for brain tumor segmentation in clinical scenarios with limited modalities. The core innovation lies in the novelty-designed EMS and CMA modules. In detail, the EMS module dynamically allocates varying attention weights to different imaging modalities across anatomical regions. Meanwhile, the CMA module establishes bidirectional cross-talk between encoder and decoder pathways through skip connections and leverages a mixed attention mechanism to refine feature representation. Additionally, the knowledge distillation framework is adopted which effectively enhances segmentation performance in scenarios with limited imaging modalities. Comprehensive evaluations on the BraTS2020 challenge benchmark dataset demonstrate EMSA-Net's su-

Table 7: Comparison of model parameters and computational complexity.

Method	RobustSeg	MFI	M2FTrans	RFNet	Ours
Parms (M)	37.58	30.91	42.87	42.30	33.53
GFLOPs (B)	846.54	999.04	209.57	204.57	333.10

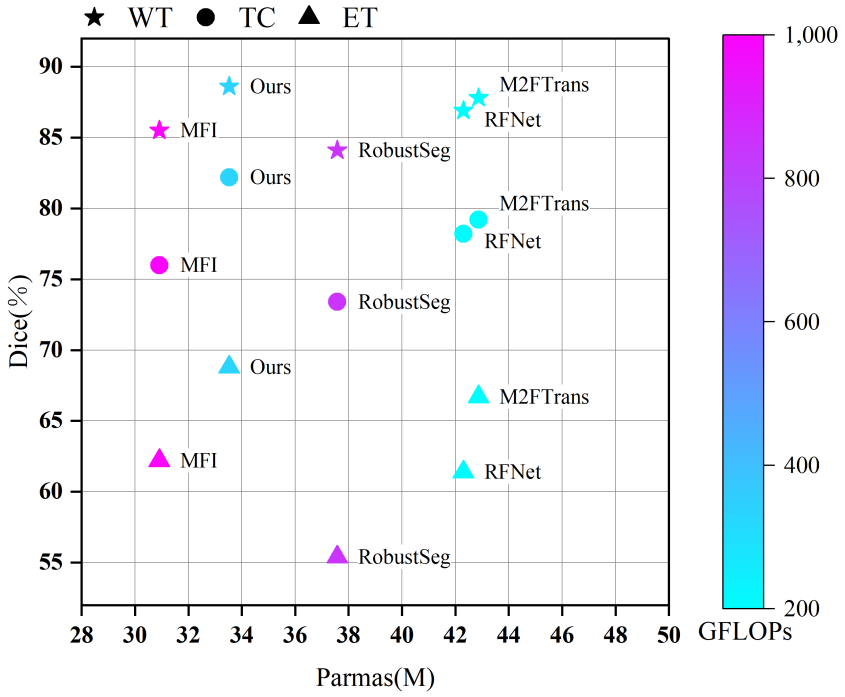


Figure 8: Different methods on model efficiency. Model parameters (i.e., Params) is measured in millions. GFLOPs is the abbreviation of floating-point operations per second measured in billions.

periority over contemporary state-of-the-art methods. Remarkably, the proposed EMSA-Net achieves consistent performance gains across all 15 multi-modality configurations, validating its robustness and generalizability. In future work, the lightweight network design will be further explored to enhance the model's efficiency and applicability in real-world clinical settings.

Declarations

Ethics approval and consent to participate

This study utilized the publicly available BraTS2020 challenge dataset. The dataset organizers have confirmed that all data were collected with patient consent and institutional review board approval. Therefore, no further ethics approval was required for this retrospective study of anonymized data.

Consent for publication

Not applicable. The data used in this study were fully anonymized before being made publicly available. The manuscript does not contain any individual person's data in any form.

Availability of data and material

The BraTS2020 dataset analyzed during the current study is publicly available from the official challenge website `https://www.med.upenn.edu/cbica/brats2020/`. The source code supporting the findings of this study is available from the corresponding author on reasonable request.

Funding

This work was supported by the Key Science and Technology Program of Henan Province (242102211042), Fundamental Research Funds for the Universities of Henan Province (NSFRF2502060), HPU Doctoral Research Fund (B2022-14), HenanPostdoctoral Foundation.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' contributions

B. Hou: Conceptualization, Funding Acquisition, Methodology, Resources, Supervision, Review.

Y. Wang: Investigation, Conceptualization, Formal Analysis, Visualization, Writing, Validation – Original Draft.

Z. Xu: Methodology, Review, Supervision.

J. Sun: Supervision.

L. Zhang: Supervision.

L. Zhao: Resources, Supervision.

B. Pu: Supervision, Review.

Y. Zhang: Supervision.

All authors have read and approved the final manuscript.

References

[1] S. Yin, Y. Wang, S. Wang, J. Chang, J. Zhu, W. Lu, Tumor segmentation in brain mr images using res-unet with multimodal fusion, in: 2025 Asia-Europe Conference on Cybersecurity, Internet of Things and Soft Computing (CITSC), 2025, pp. 28–32. doi:10.1109/CITSC64390.2025.00011.

[2] A. Verma, A. K. Yadav, Optimized U-Net Architecture with Enhanced Chimp Optimization Algorithm for Brain Tumor Segmentation, Cognitive Computation 17 (2025). doi:10.1007/s12559-025-10530-1.

[3] C. Dong, D. Sun, B. Luo, HLAE: Hierarchical Local Attention Encoder for MRI Brain Tumor Image Classification, Cognitive Computation 17 (2025). doi:10.1007/s12559-025-10419-z.

[4] M. V. Madhu Thota, B. R. Prasad, Incs: A novel brain tumor detection methodology design based on deep learning assisted intelligent neural classification strategy, in: 2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), 2024, pp. 1–5. doi:10.1109/ICSES63760.2024.10910369.

[5] J. Akonda, S. T. Alvi, C. A. A. Era, S. J. Mitu, M. N. Uddin, Mri based brain tumor detection comprising a hybrid segmentation technique, in: 2024 International Conference on Decision Aid Sciences and Applications (DASA), 2024, pp. 1–6. doi:10.1109/DASA63652.2024.10836563.

[6] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, Y. Wang, Y. Yu, Exploring task structure for brain tumor segmentation from multi-modality mr images, IEEE Transactions on Image Processing 29 (2020) 9032–9043. doi:10.1109/TIP.2020.3023609.

[7] S. Yin, Y. Wang, S. Wang, J. Chang, J. Zhu, W. Lu, Tumor segmentation in brain mr images using res-unet with multimodal fusion, in: 2025 Asia-Europe Conference on Cybersecurity, Internet of Things and Soft Computing (CITSC), 2025, pp. 28–32. doi:10.1109/CITSC64390.2025.00011.

[8] D. S, B. P S, D. R, K. K M. E., Deep learning-driven brain tumor segmentation from mri scans, in: 2025 International Conference on

Electronics and Renewable Systems (ICEARS), 2025, pp. 1982–1986. doi:10.1109/ICEARS64219.2025.10940593.

[9] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, J. Wu, Unet 3+: A full-scale connected unet for medical image segmentation, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 1055–1059. doi:10.1109/ICASSP40776.2020.9053405.

[10] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, D. Rueckert, Attention u-net: Learning where to look for the pancreas, in: Medical Imaging with Deep Learning, 2018.
URL https://openreview.net/forum?id=Skft7cijM

[11] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, M. P. Lungren, S. Zhang, L. Xing, L. Lu, A. Yuille, Y. Zhou, Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers, Medical Image Analysis 97 (2024) 103280. doi:https://doi.org/10.1016/j.media.2024.103280.
URL https://www.sciencedirect.com/science/article/pii/S1361841524002056

[12] L. Shen, W. Zhu, X. Wang, L. Xing, J. M. Pauly, B. Turkbey, S. A. Harmon, T. H. Sanford, S. Mehralivand, P. L. Choyke, B. J. Wood, D. Xu, Multi-domain image completion for random missing input data, IEEE Transactions on Medical Imaging 40 (4) (2021) 1113–1122. doi:10.1109/TMI.2020.3046444.

[13] D. Lee, J. Kim, W.-J. Moon, J. C. Ye, Collagan: Collaborative gan for missing image data imputation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2482–2491. doi:10.1109/CVPR.2019.00259.

[14] J. Shi, L. Yu, Q. Cheng, X. Yang, K.-T. Cheng, Z. Yan, Mftrans: Modality-masked fusion transformer for incomplete multi-modality brain tumor segmentation, IEEE Journal of Biomedical and Health Informatics 28 (1) (2024) 379–390. doi:10.1109/JBHI.2023.3326151.

[15] C. Chen, Q. Dou, Y. Jin, Q. Liu, P. A. Heng, Learning with privileged multimodal knowledge for unimodal segmentation, IEEE Transactions on Medical Imaging 41 (3) (2022) 621–632. doi:10.1109/TMI.2021.3119385.

[16] M. Hu, M. Maillard, Y. Zhang, T. Ciceri, G. La Barbera, I. Bloch, P. Gori, Knowledge distillation from multi-modal to mono-modal segmentation networks, in: A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, L. Joskowicz (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Springer International Publishing, Cham, 2020, pp. 772–781.

[17] Y. Ding, X. Yu, Y. Yang, Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3955–3964. doi:10.1109/ICCV48922.2021.00394.

[18] Y. Zhang, N. He, J. Yang, Y. Li, D. Wei, Y. Huang, Y. Zhang, Z. He, Y. Zheng, mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation, in: L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, S. Li (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Springer Nature Switzerland, Cham, 2022, pp. 107–117.

[19] S. R. Govind, B. Moparthi, A. Oliver, M.Kituthika, Robust brain tumor classification using transfer learning and convolutional neural networks, in: 2025 International Conference on Pervasive Computational Technologies (ICPCT), 2025, pp. 489–495. doi:10.1109/ICPCT64145.2025.10940511.

[20] W. Wang, L. Wang, X. Wang, Gba-net: A method for 3d brain tumor segmentation based on multi-scale gaussian boundary attention, in: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025, pp. 1–5. doi:10.1109/ICASSP49660.2025.10889479.

[21] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, J. Li, Transbts: Multimodal brain tumor segmentation using transformer, in: M. de Bruijne, P. C.

Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, C. Essert (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham, 2021, pp. 109–119.

[22] A. Verma, A. K. Yadav, Assessing u-net architectures for effective brain tumor segmentation in mri, in: 2025 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), 2025, pp. 1–5. doi:10.1109/SCEECS64059.2025.10940673.

[23] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241.

[24] M. Zhang, Q. Sun, Y. Han, J. Zhang, Edge-interaction mamba network for mri brain tumor segmentation, in: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025, pp. 1–5. doi:10.1109/ICASSP49660.2025.10889470.

[25] A. M. Barzili, B. Moshiri, Attention-based segmentation of brain tumors using feature level fusion and unet architecture, in: 2025 Fifth National and the First International Conference on Applied Research in Electrical Engineering (AREE), 2025, pp. 1–6. doi:10.1109/AREE63378.2025.10880261.

[26] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, O. Ronneberger, 3d u-net: Learning dense volumetric segmentation from sparse annotation, in: S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, W. Wells (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, Springer International Publishing, Cham, 2016, pp. 424–432.

[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical Vision Transformer using Shifted Windows , in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 9992–10002. doi:10.1109/ICCV48922.2021.00986.
URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00986

[28] X. L. Pham, M. H. Luu, T. van Walsum, H. S. Mai, S. Klein, N. H. Le, D. T. Chu, Cman: Cascaded multi-scale spatial channel attention-guided network for large 3d deformable registration of liver ct images, Medical Image Analysis 96 (2024) 103212. doi:https://doi.org/10.1016/j.media.2024.103212.
URL https://www.sciencedirect.com/science/article/pii/S1361841524001373

[29] M. Jaderberg, K. Simonyan, A. Zisserman, k. kavukcuoglu, Spatial transformer networks, in: C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 28, Curran Associates, Inc., 2015.

[30] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141. doi:10.1109/CVPR.2018.00745.

[31] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[32] J. Li, S. Zhu, Channel-spatial transformer for efficient image super-resolution, in: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 2685–2689. doi:10.1109/ICASSP48485.2024.10446047.

[33] Z. Li, Y. Zhang, H. Li, Y. Chai, Y. Yang, Deformation-aware and reconstruction-driven multimodal representation learning for brain tumor segmentation with missing modalities, Biomedical Signal Processing and Control 91 (2024) 106012. doi:https://doi.org/10.1016/j.bspc.2024.106012.
URL https://www.sciencedirect.com/science/article/pii/S1746809424000703

[34] C. Qiu, Y. Song, Y. Liu, Y. Zhu, K. Han, V. S. Sheng, Z. Liu, Mmmvit: Multiscale multimodal vision transformer for brain tumor segmentation with missing modalities, Biomedical Signal Processing and Control 90 (2024) 105827. doi:https://doi.org/10.1016/j.bspc.2023.105827.
URL https://www.sciencedirect.com/science/article/pii/S1746809423012600

[35] Y. Liu, L. Fan, C. Zhang, T. Zhou, Z. Xiao, L. Geng, D. Shen, Incomplete multi-modal representation learning for alzheimer's disease diagnosis, Medical Image Analysis 69 (2021) 101953. doi:https://doi.org/10.1016/j.media.2020.101953.
URL https://www.sciencedirect.com/science/article/pii/S1361841520303170

[36] C. Zhang, Y. Cui, Z. Han, J. T. Zhou, H. Fu, Q. Hu, Deep partial multi-view learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (5) (2022) 2402–2415. doi:10.1109/TPAMI.2020.3037734.

[37] A. Kebaili, J. Lapuyade-Lahorgue, P. Vera, S. Ruan, Amm-diff: Adaptive multi-modality diffusion network for missing modality imputation (01 2025). doi:10.48550.

[38] C. Li, B. C. Lovell, Uncertainty-guided cross-modality semi-supervised learning for mri segmentation, in: 2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2023, pp. 65–72. doi:10.1109/DICTA60407.2023.00018.

[39] R. Dorent, S. Joutard, M. Modat, S. Ourselin, T. Vercauteren, Hetero-modal variational encoder-decoder for joint modality completion and segmentation, in: D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, A. Khan (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Springer International Publishing, Cham, 2019, pp. 74–82.

[40] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network (03 2015).

[41] S. Wang, Z. Yan, D. Zhang, H. Wei, Z. Li, R. Li, Prototype knowledge distillation for medical segmentation with missing modality, in: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5. doi:10.1109/ICASSP49357.2023.10095014.

[42] C. Chen, Q. Dou, Y. Jin, H. Chen, J. Qin, P.-A. Heng, Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion, in: D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, A. Khan (Eds.), Medical Image Computing and Computer

Assisted Intervention – MICCAI 2019, Springer International Publishing, Cham, 2019, pp. 447–456.

[43] T. Zhou, S. Canu, P. Vera, S. Ruan, Latent correlation representation learning for brain tumor segmentation with missing mri modalities, IEEE Transactions on Image Processing 30 (2021) 4263–4274. doi:10.1109/TIP.2021.3070752.

[44] M. Havaei, N. Guizard, N. Chapados, Y. Bengio, Hemis: Hetero-modal image segmentation, in: S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, W. Wells (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, Springer International Publishing, Cham, 2016, pp. 469–477.

[45] L. Xing, M. Chen, J. Yao, Y. Zhang, Y. Wang, Pre-post interaction learning for brain tumor segmentation with missing mri modalities, in: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 1711–1715. doi:10.1109/ICASSP48485.2024.10446907.

[46] Z. Zhao, H. Yang, J. Sun, Modality-adaptive feature interaction for brain tumor segmentation with missing modalities, in: L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, S. Li (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Springer Nature Switzerland, Cham, 2022, pp. 183–192.