Scalable Swin Transformer Network for Brain Tumor Segmentation from Incomplete MRI Modalities

Dongsong Zhang[a,c], Changjian Wang[b], Tianhua Chen[c], Weidao Chen[d], Yiqing Shen[e,*]

[a] *School of Big Data and Artificial Intelligence, Xinyang College, Xinyang, 464000, Henan, China*
[b] *National Key Laboratory of Parallel and Distributed Computing, Changsha, 410073, Hunan, China*
[c] *School of Computing and Engineering, University of Huddersfield, Huddersfield, HD13DH, UK*
[d] *Beijing Infervision Technology Co., Ltd., Beijing, 100020, China*
[e] *Department of Computer Science, Johns Hopkins University, Baltimore, 21218, Maryland, USA*

Abstract

Background: Deep learning methods have shown great potential in processing multi-modal Magnetic Resonance Imaging (MRI) data, enabling improved accuracy in brain tumor segmentation. However, the performance of these methods can suffer when dealing with incomplete modalities, which is a common issue in clinical practice. Existing solutions, such as missing modality synthesis, knowledge distillation, and architecture-based methods, suffer from drawbacks such as long training times, high model complexity, and poor scalability.

Method: This paper proposes IMS^{2}Trans, a novel lightweight scalable Swin Transformer network by utilizing a single encoder to extract latent feature maps from all available modalities. This unified feature extraction process enables efficient information sharing and fusion among the modalities, resulting in efficiency without compromising segmentation performance even in the presence of missing modalities.

Results: Two datasets, BraTS 2018 and BraTS 2020, containing incom-

*Corresponding Author.
Email addresses: D.Zhang@hud.ac.uk (Dongsong Zhang), T.Chen@hud.ac.uk (Tianhua Chen), yshen92@jhu.edu (Yiqing Shen)

plete modalities for brain tumor segmentation are evaluated against popular benchmarks. On the BraTS 2018 dataset, our model achieved higher average Dice similarity coefficient (DSC) scores for the whole tumor, tumor core, and enhancing tumor regions (86.57, 75.67, and 58.28, respectively), in comparison with a state-of-the-art model, i.e. mmFormer (86.45, 75.51, and 57.79, respectively). Similarly, on the BraTS 2020 dataset, our model scored higher DSC scores in these three brain tumor regions (87.33, 79.09, and 62.11, respectively) compared to mmFormer (86.17, 78.34, and 60.36, respectively). We also conducted a Wilcoxon test on the experimental results, and the generated p-value confirmed that our model's performance was statistically significant. Moreover, our model exhibits significantly reduced complexity with only 4.47M parameters, 121.89G FLOPs, and a model size of 77.13MB, whereas mmFormer comprises 34.96M parameters, 265.79G FLOPs, and a model size of 559.74MB. These indicate our model, being light-weighted with significantly reduced parameters, is still able to achieve better performance than a state-of-the-art model.

Conclusion: By leveraging a single encoder for processing the available modalities, IMS^{2}Trans offers notable scalability advantages over methods that rely on multiple encoders. This streamlined approach eliminates the need for maintaining separate encoders for each modality, resulting in a lightweight and scalable network architecture. The source code of IMS^{2}Trans and the associated weights are both publicly available at `https://github.com/hudscomdz/IMS2Trans`.

Keywords: Incomplete Modality, Brain Tumor Segmentation, Transformer
PACS: 0000, 1111
2000 MSC: 0000, 1111

1. Introduction

Magnetic resonance imaging (MRI) is a widely-used non-invasive imaging technique for clinical assessment and therapy planning for tumors in soft tissues such as the brain [1]. To obtain a comprehensive characterization of the anatomy, MRIs are typically acquired with different contrast, resulting in multi-modal MRI scans[1] [5]. Common MRI modalities [5] include T1-weighted (T1w), contrast-enhanced T1-weighted (T1c), T2-weighted (T2w),

[1]In some literature [2, 3, 4], 'modality' is also termed as 'sequence'.

Fluid Attenuation Inversion Recovery (FLAIR), Magnetization Prepared RAPid Gradient Echo (MP-RAGE), and Proton Density (PD-w). Considering that invasive growth brain tumors are usually fused with brain soft tissues, it is difficult to accurately segment tumor structures using single-modality MRI images. Instead, by providing complementary information, the availability of multi-modal brain MRI data can improve the accuracy of lesion identification, and disease diagnosis for both human and computer-aided diagnosis (CAD) systems such as deep learning models. Consequently, in terms of brain tumor segmentation from MRI, various feature fusion strategies have developed upon convolutional neural network (CNN) [6] or Transformer [7]. Li et al. [8] proposed a dual X-Net codec structure combining the characteristics of CNN and Transformer, which extracts local and global features simultaneously through convolution subsampling and Transformer encoders and then reconstructs the input image itself through variational autoencoder branches in the decoding stage. The experiment shows that X-Net can realize the organic combination of Transformer and CNN. Xu et al. [9] proposed a hybrid feature extraction network, which fully integrates the features extracted by CNN and Transformer to enhance the segmentation performance of brain tumor medical images. Zhu et al. [10] proposed a brain tumor segmentation method that integrates multi-modal MRI information, which combines deep semantic and edge information fusion, using Swin Transformer for feature extraction, CNN-based edge detection module, and multi-feature inference block based on graph convolution. To achieve real-time medical image segmentation, He et al. [11] proposed a cloud-based method based on multi-feature extraction and interactive fusion. The method uses Transformer and CNN to extract global and local features, respectively. The interactive fusion focus module improves segmentation accuracy. Lu et al. [12] proposed a 3D multi-scale Ghost convolution neural network (GMetaNet) with an auxiliary MetaFormer decoding path, which combines local modeling of CNN with remote representation of Transformer to achieve efficient semantic information extraction of multi-modal brain tumor MRI images. To address the issue of neural networks using too many parameters and being difficult to deploy, Liu et al. [13] proposed a lightweight 3D brain tumor image segmentation method with hierarchical decoupled convolutions that reduces the number of parameters, which also uses an attention mechanism in the output layer to improve segmentation accuracy.

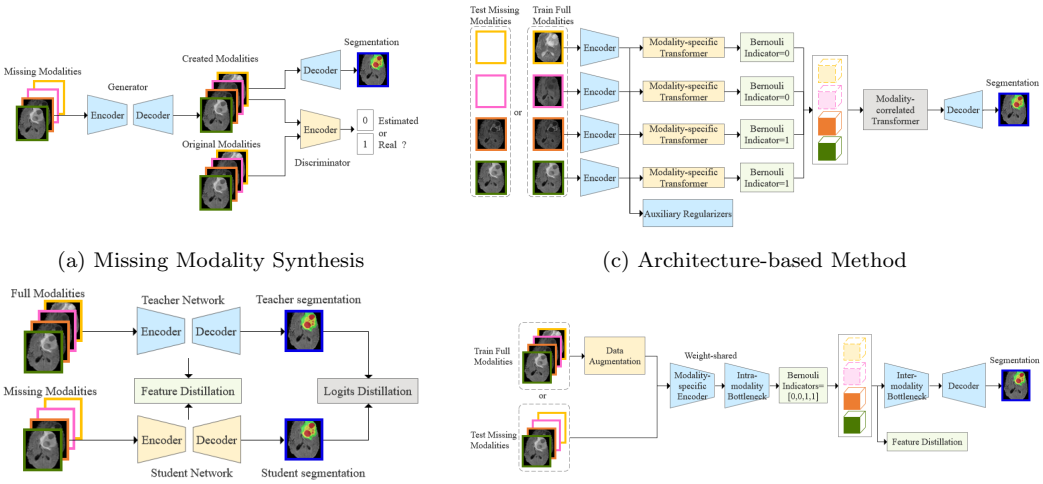
However, certain factors may lead to missing MRI modalities [14][2], while most of the existing multi-modal deep learning methods are not applicable to address this issue. For example, one possible reason for missing MRI modalities is that patients may fail to comply with instructions from radiologists or clinicians [15], which can compromise the quality of the scans of specific modalities. Another factor is the acquisition time constraints during scanning, due to the cost and considerations of patient comfort [16], which may prevent the collection of all required MRI modalities. Additionally, body movements during the scan can lead to artifacts and unusable low-quality images [17], resulting in the loss of certain modalities. Finally, the change of MRI imaging protocols can also contribute to unaligned or the absence of MRI sequences [18].

To address the missing modalities in multiple-parametric MRI analysis, which is a common issue in brain tumor segmentation, several remedies have been proposed. They can be broadly classified into three categories [5, 19]. The first category, as depicted in Figure 1(a), employs generative models such as Generative Adversarial Network (GAN) [20], Diffusion Models [21] to synthesize the missing modalities from observed MRI modalities as data preprocessing. However, this approach has the drawback of requiring an additional model to be trained before downstream analysis, leading to longer training and execution times, as well as the accumulation of errors. As illustrated in Figure 1(b), the second category involves using knowledge distillation to extract feature representations from a teacher network trained with full modalities to a student network specifically tailored for missing modalities [22, 23, 24, 25, 26]. Yet, distillation-based methods require a series of students to tackle each condition of missing modalities, leading to a huge computation of spatial and time costs. The third category utilizes a single network that can directly handle any conditions of missing modalities for particular downstream tasks [27, 28, 29, 30, 31, 32, 33, 34, 35], as shown in Figure 1(c). However, this line of approaches encounters limitations such as large network parameters, slow training speed, and poor scalability, as they require one encoder for each modality.

Being able to accurately segment brain tumors from MRI scans is crucial for diagnosis, treatment planning, and assessing response to therapy. How-

[2]We follow the literature by using 'missing modalities' and 'incomplete modality' interchangeably.

ever, it is common to encounter incomplete modalities in clinical practice due to various factors such as patient non-compliance, time constraints, artifacts, and changes in protocols [14, 15, 16, 17, 18]. Existing multi-modal deep learning methods suffer performance declines when confronted with missing modalities [5, 19]. To address this critical issue, we propose a novel scalable Swin Transformer network specifically engineered to maintain segmentation performance even when MRI modalities are incomplete or absent. Our method offers a lightweight and efficient solution that requires significantly fewer parameters compared to previous methods that rely on multiple modality-specific encoders. By using a single shared-weight encoder for feature extraction coupled with our proposed data augmentation and distillation techniques, our network provides an important advancement that can effectively and efficiently handle missing modalities while ensuring accurate brain tumor segmentation.



(a) Missing Modality Synthesis

(c) Architecture-based Method

(b) Knowledge Distillation

(d) Our Proposed Method

Figure 1: Overview of the four remedies for missing MRI modalities.

To narrow the gap, we propose Incomplete Modalities Scalable Swin Transformer (IMS^{2}Trans), a novel lightweight and scalable network architecture for incomplete MRI multi-modal brain tumor segmentation. As illustrated in Figure 1(d), our method employs a single encoder with shared weights to extract features from all the observed modalities along with a feature distillation scheme to impose consistency regularization among modalities, before being finally aggregated to obtain the segmentation result via

the decoder. The major contributions are four-fold. (1) To the best of our knowledge, we first propose a scalable Swin Transformer [36] as the only share-weighted encoder for the incomplete MRI modalities of brain tumor segmentation. Specifically, all the available observed modalities are input to an encoder with shared weights to reduce the number of parameters and increase efficiency. Correspondingly, a novel modality token strategy is designed to specify the difference between input modalities. (2) We introduce a Swin-like lightweight MLP bottleneck [37] that not only reduces model parameters but also obtains better feature maps of intra-modalities and inter-modalities. (3) We also design a new feature distillation regularization based on contrastive learning [38] to improve the interchangeability and consistency across different modalities. (4) We propose a novel 3D multimodal version based on the CutMix [39] data augmentation strategy specifically for multi-modal MRI data further to enhance the model robustness against the missing modalities.

2. Related Work

In this section, we review the current solutions for missing modalities issues in MRI analysis, and highlight how our method differs from them.

2.1. Missing Modality Synthesis

The field of multiple-parametric MRI analysis has seen the emergence of generative methods in addressing the issues of missing modalities by synthesizing the missing data from observed modalities. These methods can be classified into three categories: (i) one-to-one synthesis methods [40, 41], (ii) many-to-one synthesis methods [42, 21], (iii) unified synthesis methods [43, 44, 45].

One-to-one synthesis approaches generate one target contrast from a single input observed modality. For instance, a multi-contrast MRI synthesis method that uses a conditional GAN [46] was utilized to synthesize missing target contrasts for T1w and T2w MRIs in an unimodal input setting [40]. Additionally, Ea-GAN [41] was proposed to synthesize missing MR images in a one-to-one setting by integrating critical textural edge information from the input modality to the edge maps to further enhance the synthesized image quality. However, the one-to-one synthesis method has a significant drawback in that it can be computationally expensive due to the fact that each synthesis condition requires training a separate model. Therefore, the

more modalities required for analysis, the greater the computational burden becomes.

Many-to-one synthesis approaches, on the other hand, learn a mapping from multiple available modalities to one target contrast. For instance, MUSTGAN has been proposed for the joint one-to-one and many-to-one synthesis [42]. Recently, the diffusion model (DM) [47] has been shown to perform better than the GAN model in many-to-one MRI synthesis. For example, a DM-based many-to-one MRI synthesis model [21] has been proposed, which generates missing modalities from multiple available modalities. While many-to-one synthesis methods can capture shared features from multiple observed modalities, they struggle to effectively obtain unique and complementary features from specific modalities, which ultimately limits their overall synthesis performance.

Unified synthesis approaches can generate one or more target contrasts from any available modalities, as exemplified by MM-GAN [43]. However, as the number of missing modalities increases, the performance of these approaches may degrade. Furthermore, most of these methods either use multiple encoding and decoding streams [44] or a complicated synthetic network such as Hyper-GAE [45], which increases memory consumption and impairs the scalability of the model, making it challenging to handle a larger number of modalities.

In other words, generative methods suffer from two main shortcomings. Firstly, it is computationally expensive and requires significant network training time, especially with numerous missing modalities. Secondly, missing modalities generated from scratch may have defects, such as failing to synthesize critical features of the original missing modality. In contrast, our proposed method has fewer parameters, consumes less memory, and can scale better to accommodate multiple modalities.

2.2. Knowledge Distillation

Knowledge distillation based approach aims to recover missing modality features by transferring the extracted features from the full modality path of the teacher network to the student network for the missing modality [22, 23, 24, 25, 26, 48, 49, 50, 51]. For example, KDD-Net trains the teacher model on all available modalities and then is distilled to the student model using the feature representations from available modalities [22]. Similarly, HAD-Net, a hierarchical adversarial distillation network, includes three

main components: the teacher network, the student network, and the hierarchical discriminator [23]. The teacher network in HAD-Net is trained on all complete modalities, and the student network is trained on all modalities except T1c MRIs. The hierarchical discriminator can bridge the feature gap by distinguishing the segmentation results of the student from the teacher and mapping the multi-scale feature representations into a common latent representation space. These methods only focus on recovering the missing T1c MRI, while others use more collaborative approaches. ACN uses an adversarial co-training network of both full and missing modalities to recover missing modalities [24], while D^2-Net leverages a dual disentanglement network to segment brain tumors with missing modalities and identify relationships between them [25]. Although knowledge distillation can improve feature learning in the student network, it is still challenging for the student model to obtain important feature representations from all modalities.

2.3. Architecture-based Method

Inspired by the fact that different MRI modalities represent the same brain tumor region [8, 9, 10, 11], a research line has emerged to mine correlations between the feature distributions of different modalities [28, 31, 32]. One approach has been to use networks that encode each modality individually and provide them with a correlation block. However, this method may not recover lost information if the number of available modalities is insufficient. To address the above-mentioned limitation, recent methods [30, 33, 34, 35] have proposed using attention mechanisms for brain tumor segmentation tasks with missing modalities. For example, Ding *et al.* [30] introduced a novel region-aware fusion module that divides multi-modal features into different regions using a trained probability map and then applies modal-wise attention to adjust features from available modalities. Additionally, Zhang *et al.* [33] proposed mmFormer that combines Transformer blocks and convolutional encoders to build local and global information within each modality and long-range correlations across modalities, representing the first attempt to achieve this using Transformer blocks. Zhou *et al.* [34] suggested a new multi-modality feature fusion network that uses a self-attention mechanism to learn non-local structures in images across multiple modalities, and a multi-scale fusion module to capture feature information in multi-modality spatial contexts, as well as a spatially consistent underlying feature learning module to learn potential multi-modality correlations. Furthermore, Konwer *et al.* [35] proposed a new method to address brain tumor image segmentation

with incomplete modalities by introducing an auxiliary adversarial learning strategy to supervise the representation of missing modality features during meta-training of partial modality data and meta-testing of limited full modality subjects.

While those methods have made significant improvements to MRI analysis for missing modalities, they still face challenges in scenarios where more than one modality is missing and have higher memory consumption due to the large number of parameters arising from an equal number of encoders to the number of modalities. In contrast, the proposed method has a unique advantage in that it reuses its encoder to encode multiple modalities, and the weights of the encoder are shared. Our approach significantly reduces the number of parameters the network requires while maintaining performance, making it more memory-efficient and scalable to handle multiple modalities.

3. Methodology

3.1. Overview

In this section, we elaborate on our novel IMS^{2}Trans network, specifically engineered for the segmentation of brain tumors in MRI scans with arbitrary MRI modalities missing. A schematic of this network is provided in Figure 2. At the core of our network resides a scalable shared-weight encoder that leverages the Swin Transformer block [36] with modality token to capture both local and global context. This shared-weight design enables the efficient encoding of multiple modalities using a single encoder, thus optimizing computational resources and reducing overall model complexity. To further enhance our encoder, we introduce a lightweight Shifted Multi-Layer Perceptron (Shifted MLP) [37] coupled with a masking bottleneck. This combination is designed to balance computational complexity with high-level accuracy, particularly in dealing with missing modalities. We also implement a feature distillation strategy between individual modalities and the entire set of modalities, by comparing the features of each modality with the averaged features computed across all modalities to ensure a more comprehensive and accurate representation of features in a missing modality circumstance. Lastly, to boost the performance and adaptability of our network, especially in the context of missing modalities, we adopt a unique data augmentation technique: the 3D multimodal CutMix (3DMM-CutMix)[39]. This approach strengthens the network's resilience and adaptability, setting the stage for superior performance under varying conditions.

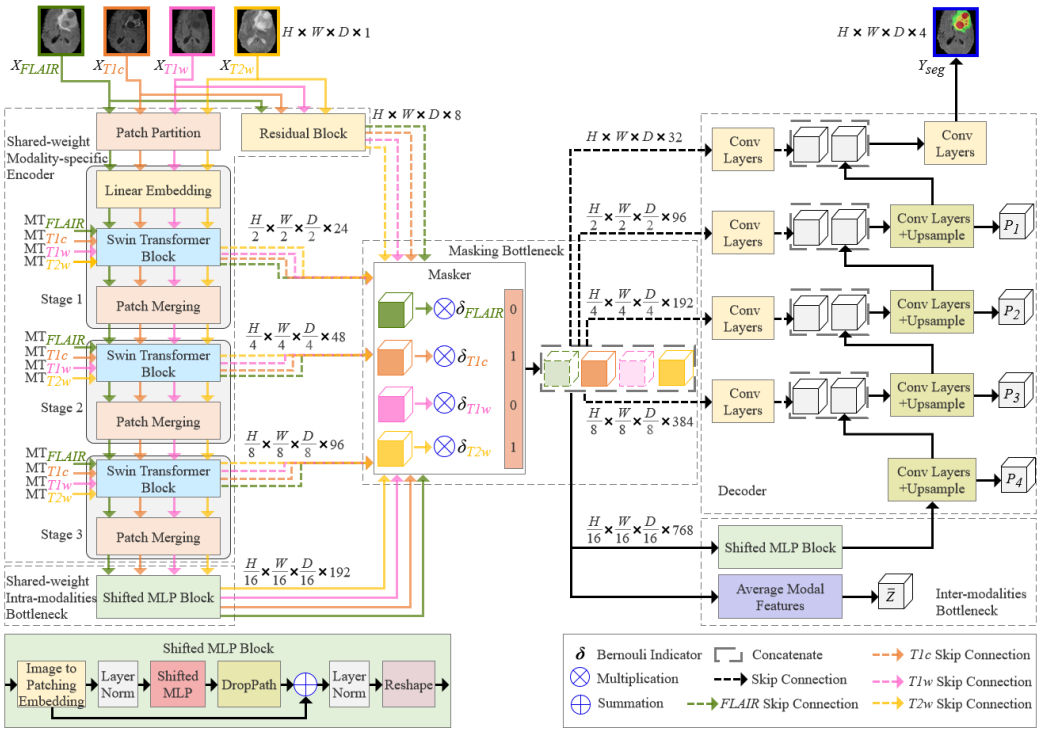


Figure 2: Overview of the proposed IMS^{2}Trans network. It comprises a scalable shared-weight modality-specific encoder, intra-, and inter-modalities shifted MLP bottlenecks, a masking bottleneck, a multimodal feature distillation module, and a convolutional decoder. Skip connections are strategically applied to the encoder, masker, and decoder to enhance feature learning. For clarity in display, the 3DMM-CutMix data augmentation technique applied to the multimodal input image is not depicted.

3.2. Scalable Shared-weighted Encoder

The scalable shared-weighted encoder is designed to process 3D MRI image modalities by a single encoder to reduce the number of parameters and improve efficiency. Concurrently, each modality retains its distinct characteristics, which are ensured by a unique, learnable modality token. Each input MRI modality image, denoted as $X_i \in \mathbb{R}^{H\times W\times D\times 1}$, where $i \in$ {FLAIR, T1C, T1w, T2w} denotes the corresponding modality, is first divided into non-overlapping patches through patch partition operator. Here, H, W, and D signify the height, width, and number of slices in the modality image respectively. The patches, each with a dimension of $2 \times 2 \times 2$, result in a feature dimension of 8. These patch tokens are then transformed into an embedding space of dimension $C = 24$ using a learnable linear layer. In parallel, the

input X_i is directed through a residual block, resulting in an 8-dimensional token.

For each input modality image, a corresponding token is produced, thus there are as many such tokens as there are the number of input modalities. These tokens are then routed to the masking bottleneck, concatenated, and subsequently passed into the decoder. Post this, the embedded feature map of input X_i is fed into the encoder, which consists of three layers of the swin transformer. Each of these layers encompasses two transformer blocks with modality tokens and is succeeded by a patch merging module. Notably, the modality token denoted as MT_i, $i \in \{\text{FLAIR, T1c, T1w, T2w}\}$, representing the embedded feature of each modality, plays a crucial role in retaining the distinctive characteristics of each modality. The size of MT_i in each block is the same as that of the input X_i. The patch merging module reduces the feature dimensions, thereby promoting efficient computation and enabling hierarchical feature extraction. Moreover, the patch merging module combines patches of resolution $2 \times 2 \times 2$ and concatenates them, forming a $4C$-dimensional feature embedding. This is further condensed to a $2C$-dimensional feature size by another linear layer. As a result, the resolutions after the first, second, and third swin transformer layers become $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}$, $\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}$, and $\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}$ respectively, while the corresponding channel number of the embedding space C incrementally increases to 48, 96, and 192 respectively. To further enhance computational efficiency and maintain critical feature characteristics, a convolutional layer is tactically positioned between each pair of consecutive swin transformer layers, thereby facilitating the down-sampling of the feature map.

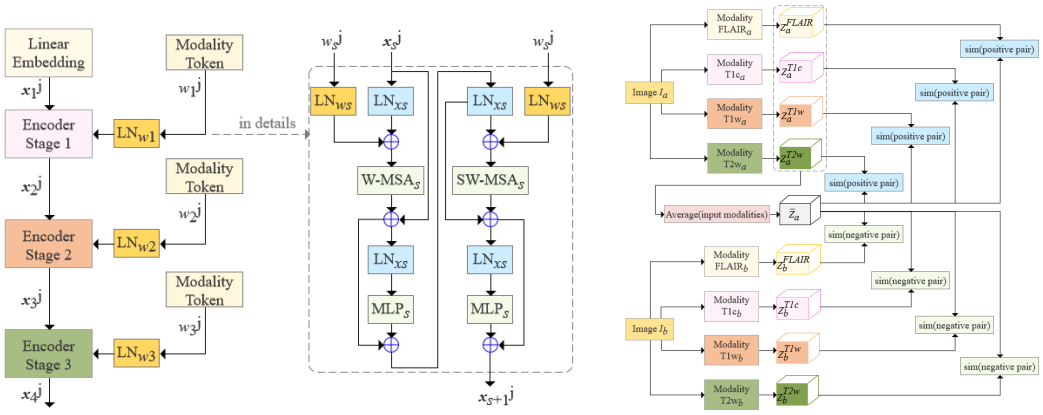
3.3. Swin Transformer Block with Modality Token

Each input modality image comes with its own information about which modality it is, from which a corresponding token can be designed to represent the embedded features of each modality. As a result, the encoder's architecture, featuring the swin transformer block with the novel modality token, is illustrated in Figure 3(a). Each stage of the encoder in this figure corresponds one-to-one to each stage of the shared-weight modal-specific encoder in Figure 2. Following the design of swin transformer [36], the swin transformer block with modality token primarily consists of two consecutive swin transformers, each comprising layer normalization (LN) modules, a multi-head self-attention module, and a multi-layer perceptron (MLP) with GELU nonlinear activation. Each of these components is linked via a residual

connection with DropPath [52]. The first and second consecutive swin transformers employ window-based (W-MSA) and shifted window-based (SW-MSA) multi-head self-attention modules, respectively. These modules are essentially variations of the regular and window-partitioning multi-head self-attention modules. According to Figure 2, the modality tokens of the four input mode images are MT_{FLAIR}, MT_{T1c}, MT_{T1w}, and MT_{T2w}. Consider x_i^j and w_i^j to represent the feature token of the jth modality and the corresponding modality token feeding the swin transformer block at each encoder stage, respectively. As per the window division mechanism, the swin transformer block with a modality token can be formally expressed as follows:

$$\begin{cases} \hat{x}_s^j &= \text{W-MSA}(\text{LN}(x_s^j) + \text{LN}(w_s^j)) + x_s^j, \\ x_{s'}^j &= \text{MLP}(\text{LN}(\hat{x}_s^j)) + \hat{x}_s^j, \\ \hat{x}_{s+1}^j &= \text{SW-MSA}(\text{LN}(x_{s'}^j) + \text{LN}(w_s^j)) + x_{s'}^j, \\ x_{s+1}^j &= \text{MLP}(\text{LN}(\hat{x}_{s+1}^j)) + \hat{x}_{s+1}^j, \end{cases} \tag{1}$$

where $\hat{x}_s^j$ and $\hat{x}_{s+1}^j$ denote the outputs from the W-MSA and SW-MSA modules respectively, whereas $x_{s'}^j$ and x_{s+1}^j represent the outputs of the MLP module in the first and second transformers, respectively.



(a) Modality Token (b) Feature Distillation Regularization

Figure 3: The semantic illustration of the modality token and feature distillation regularization.

3.4. Shifted MLP and Masking Bottleneck

Before being fed in the Shifted MLP bottleneck, features with respect to each modality extracted by the swin transformer encoder are first encoded

into tokens separately. These tokens are then passed onto the Shifted MLP module, as depicted in Figure 2. They are first performed an axis shift across the width of the tokens, where the locality of the token is integrated into the global axial attention computation, promoting more spatially aware features. Subsequently, tokens are funneled through a parameter-efficient depth-wise convolutional layer (DWConv) [53], after which they pass through a GELU activation layer [54], thus adding a level of learnable non-linearity. The tokens are then channeled to the second multilayer perceptron following an additional axis shift, this time along the height. The final step within the Shifted MLP bottleneck involves applying residual concatenation to append the original tokens as residuals.

Residing after the Shifted MLP bottleneck, the masking bottleneck aims to enhance the robustness of the missing modalities and where the construction of long-distance dependencies between different modalities is required. Specifically, it achieves this by generating a novel multi-modal token through the concatenation of feature maps from various modalities corresponding to the same image. These feature maps originate either from the skip connection of the encoder or the shifted MLP bottleneck. This operation is formally defined as:

$$T_{token} = [\delta_{FLAIR} \cdot T^{FLAIR}, \delta_{T1c} \cdot T^{T1c}, \delta_{T1w} \cdot T^{T1w}, \delta_{T2w} \cdot T^{T2w}], \qquad (2)$$

where $\delta_m \in \{0, 1\}$functions as a Bernoulli indicator, designed to impart robustness during the construction of long-range dependencies between different modalities, even when some modalities are missing. This form of modality-level dropout is randomly enacted during training by assigning δ_m a value of 0. Following the design in [33], the Bernoulli factor assumes a value of either 0 or 1. If δ_m equals 0, it implies the corresponding input modality is missing. Conversely, a δ_m value of 1 signifies the availability of the modality. During the training stage, multi-modal tokens for missing modalities are produced by assigning this Bernoulli factor a value of 0, effectively leading to multiplication with a zero vector.

3.5. Decoder Mechanism

The primary of the decoder is to effectively reconstruct the spatial resolution of the consolidated latent space back into the original image space. The implementation of the decoder primarily relies on convolutional neural networks, comprising a series of 3D residual convolutions and upsampling operators [33]. Specifically, the output feature map generated from the Shifted

MLP bottleneck is supplied as the decoder's input. In parallel, the decoder receives skip connections from the masking bottleneck. This dual-input approach enables the preservation of low-level details, thereby facilitating a more precise segmentation. Features derived from different modalities at varying stages of the encoder, once processed through the masking bottleneck, are concatenated and used as inputs to the convolutional layer in the decoder as skip connection features. This strategy serves a dual purpose: Firstly, it compels the decoder to generate accurate segmentation results based on low-level feature maps at each stage of the decoder. Secondly, it equips the model with the resilience to continue delivering accurate segmentation, even when confronted with the presence of missing modalities.

3.6. Feature Distillation

We innovatively integrate a feature distillation regularization based on the contrastive learning [38], which ensures uniformity among the features obtained from different modalities through the masking bottleneck, as visualized in Figure 3(b). More specifically, we generate an averaged feature map $\bar{Z}_a$ by computing the mean of the feature maps from four separate modalities of the same image for a given case, X_a: Z_a^{FLAIR}, Z_a^{T1c}, Z_a^{T1w}, Z_a^{T2w}, where Z represents the latent feature representation, the superscript denotes the corresponding modality. As the averaging process preserves the semantic essence of the image, $\bar{Z}_a$ serves as a guiding element for each individually extracted feature from each modality. The central objective is to maximize the similarity between $\bar{Z}_a$ and each of the feature maps ($Z_a^{FLAIR}, Z_a^{T1c}, Z_a^{T1w}, Z_a^{T2w}$), a measurement facilitated through the use of the cosine similarity. Conversely, feature maps drawn from different input MRI images, for instance, X_b, should present a significant divergence from $\bar{Z}_a$. In the training phase, we form positive sample pairs by combining the four input modalities from the same original MRI image with the averaged image representation, yielding at least four positive sample pairs. We generate negative samples by pairing feature maps from different input MRI images within the same batch with the averaged image representation, thereby resulting in a minimum of four negative sample pairs.

Consequently, we can formulate the feature distillation loss L_{FDC} as:

$$L_{FDC} = -\frac{1}{N}\sum_{j=1}^{N} \log P(Z_i^j | \bar{Z}_i). \tag{3}$$

Here, $P(Z_i^j|\bar{Z}_i)$ is defined as the normalized temperature-scaled softmax cosine similarity [55, 38], as follows:

$$P(Z_i^j|\bar{Z}_i) = \frac{e^{sim(Z_i^j,\bar{Z}_i)/\tau}}{\sum_{k=1}^{2N} e^{sim(Z^k,\bar{Z}_i)/\tau}}. \tag{4}$$

In these equations, Z_i^j represents the feature map of a singular modality from the MRI image X_i. $\bar{Z}_i$ is a feature map created by averaging the four modalities from the same image X_i. Z^k denotes the feature map of all samples related to $\bar{Z}_i$, including negative samples from distinct input images. The function $sim(\cdot,\cdot)$ symbolizes the cosine similarity, i.e., the dot product post-normalization. τ is the temperature, and N signifies the number of positive samples connected to $\bar{Z}_i$. It's important to note that both Z_i^j and $\bar{Z}_i$ originate from the same input MRI image, though they provide divergent feature representations of the said image.

3.7. 3DMM-CutMix

To further improve the training efficiency and performance, we propose a data augmentation strategy 3DMM-CutMix, which aims to mix two examples by replacing image regions with patches of another training image and interpolating image labels, allowing the model to focus not only on the most discriminative parts of the image but also on the entire image. The proposed 3DMM-CutMix augmentation method is designed to create a pair of new training samples, $(\tilde{X}_a^j, \tilde{G}_a)$ and $(\tilde{X}_b^j, \tilde{G}_b)$, by blending two training samples, (X_a^j, G_a) and (X_b^j, G_b). The model is then trained with this newly generated training sample set. The composition operation can be formalized as follows:

$$\begin{cases} \tilde{X}_a^j &= \mathbf{M} \odot X_a^j + (\mathbf{1} - \mathbf{M}) \odot X_b^j, \\ \tilde{G}_a &= \lambda \cdot G_a \ + \ (\mathbf{1} - \lambda) \cdot \bar{G}_a, \\ \bar{G}_a &= G_b, \\ \tilde{X}_b^j &= \mathbf{M} \odot X_b^j + (\mathbf{1} - \mathbf{M}) \odot X_a^j, \\ \tilde{G}_b &= \lambda \cdot G_b \ + \ (\mathbf{1} - \lambda) \cdot \bar{G}_b, \\ \bar{G}_b &= G_a, \end{cases} \tag{5}$$

where $\mathbf{M} \in 0, 1^{W \times H \times D}$ serves as a binary mask for the image and the label, where 0 and 1 in $\mathbf{M}$ indicate which regions in the two images are dropped or retained. The symbol $\mathbf{1}$ denotes a binary mask filled entirely with ones.

The symbol $\odot$ represents element-wise multiplication between two vectors. Mirroring the CutMix approach [39], we use λ to denote the combination ratio between two data samples. This is determined by a beta distribution, specifically $\mathbf{Beta}[\alpha, \alpha]$. As we have set the parameter α to 1, λ effectively follows a uniform distribution between 0 to 1, i.e., $U[0, 1]$. Note that the label $\tilde{G}$ in the new training example pair, as a combination of the labels G, $\bar{G}$, and λ, does not have to be actually generated, which helps our network to focus on real natural images.

The binary mask $\mathbf{M}$ with respect to the volumetric data is obtained by applying a 3D bounding box $\mathbf{B}$=($c_w, c_h, c_d, x_w, x_h, x_d$) to the cropped regions of two training MRI input images X_a^j and X_b^j within the jth modality. The binary mask $\mathbf{M}$ is constructed such that any position within $\mathbf{B}$ is set to 0, and to 1 otherwise. Effectively, the region $\mathbf{B}$ within X_a^j is excised and replaced with the corresponding cropped region $\mathbf{B}$ from X_b^j, and reciprocally, the region $\mathbf{B}$ in X_b^j is replaced by the region $\mathbf{B}$ extracted from X_a^j. Similarly, the region $\mathbf{B}$ in Y_a and region $\mathbf{B}$ in Y_b undergo the same swapping operation. Formally, $\mathbf{B}$ is determined based on the 3D image coordinates and can be described by six parameters: 3D cuboid size (c_w, c_h, c_d) and 3D center location (x_w, x_h, x_d). The coordinates of the 3D bounding box are uniformly sampled according to the following scheme:

$$\begin{aligned} &c_w = W\sqrt[3]{1-\lambda},\ c_h = H\sqrt[3]{1-\lambda},\ c_d = D\sqrt[3]{1-\lambda}, \\ &x_w \sim U[0, W], \quad x_h \sim U[0, H], \quad x_d \sim U[0, D]. \end{aligned} \tag{6}$$

Note that the scale of the cropped area is consistently maintained across all three dimensions, i.e., $\lambda = 1 - \frac{c_w c_h c_d}{WHD}$.

During each training iteration, a pair of 3DMM-CutMix samples $(\tilde{X}_a^j, \tilde{G}_a)$ and $(\tilde{X}_b^j, \tilde{G}_b)$ is produced by merging two randomly selected training samples from a mini-batch, as per the formulas provided in Eq. (5) and Eq. (6).

The implementation of our 3DMM-CutMix is detailed in Algorithm 1, where n, M, C, and K denote the batch size, the count of input modalities, the channel size of the input image, and the total segmentation classes respectively. At every training iteration, we extract a minimum batch of data (X, G) from the training set. This batch has been randomly cropped to ensure uniformity in the size of all input images. The main procedure of 3DMM-CutMix, encapsulated between lines 3 and 13, is straightforward to implement. 3DMM-CutMix generates new data $(\tilde{X}, \bar{G})$ by randomizing the order of the minibatch input images and labels along the first axis of

the tensors. Then, we sample the mixing ratio λ and the 3D bounding box $\mathbf{B}$=$(c_w, c_h, c_d, x_w, x_h, x_d)$ and the resulting cropping region $(w1, w2, h1, h2, d1, d2)$ as detailed from line 4 to 11. Subsequently, we mix the input image X and $\tilde{X}$ by replacing the cropped region of X with that of $\tilde{X}$. Note that the target labels G and $\bar{G}$ are not mixed and still denote the target labels of the two samples before being mixed, respectively. Furthermore, we adjust the value of λ to precisely match the pixel ratio. The augmented input data X is then fed into the model to obtain the predicted segmentation output Y_{seg}. The final step involves computing the loss between the predicted Y_{seg} and the target labels G and $\bar{G}$, followed by deriving the loss L_{seg} via a weighted summation based on the λ value in next section for details.

Algorithm 1 Pseudo Code for 3DMM-CutMix

Input: Random cropped training set $\{(X_i^j, G_i)\}_{i=1}^{N}{}_{j=1}^{M}$

1: **for** each_iteration **do**
2: Get a minimum batch of data $(X, G) = \{(X_i^j, G_i)\}_{i=1}^{n}{}_{j=1}^{M}$
▷ X is n×C×W×H×D size tensor, G is n×K×W×H×D size tensor.
3: $\tilde{X}, \bar{G}$ = ShuffleMinibatch(X, G) ▷ 3DMM-CutMix begins.
4: $\lambda = \mathbf{Beta}[1, 1]$
5: $c_w = W\sqrt[3]{1-\lambda}$
6: $c_h = H\sqrt[3]{1-\lambda}$
7: $c_d = D\sqrt[3]{1-\lambda}$
8: $x_w = U[0, W]$
9: $x_h = U[0, H]$
10: $x_d = U[0, D]$
11: $w1, w2, h1, h2, d1, d2$ = GetCoordinate$(c_w, x_w, c_h, x_h, c_d, x_d)$
12: $X[:, :, w1{:}w2, h1{:}h2, d1{:}d2] = \tilde{X}[:, :, w1{:}w2, h1{:}h2, d1{:}d2]$
13: $\lambda = 1 - (w2 - w1) * (h2 - h1) * (d2 - d1)/(W * H * D)$
▷ 3DMM-CutMix ends.
14: Y_{seg} = ModelForward(X)
15: Loss = ComputeLoss$(Y_{seg}, G, \bar{G}, \lambda))$
16: Update model
17: **end for**

3.8. Overall Loss

The overall loss function L_{total} is a composite of several individual loss calculations. These include L_{FDC}, corresponding to the feature distillation

method, $L_{decoder}$, resulting from the low-level feature map disparities in the decoder, and L_{seg}, which pertains to the final segmentation output. The detailed explanations and formulations for the loss function L_{FDC} are previously provided with Eq. (3).

To guide the network towards predictions more congruous with the real segmentation ground truth, we simultaneously employ Dice Similarity Coefficient (DSC) [56] and Weighted Cross Entropy (WCE)[3] [60] as metrics during the calculation of losses $L_{decoder}$ and L_{seg}. DSC, taking values between 0 and 1, evaluates the similarity between two images. It does so by determining the proportion of twice the number of intersecting voxels to the aggregate number of voxels in the prediction (Y) and the ground truth (G). This calculation is succinctly defined as follows:

$$L_{Dice}(G_{i,k}, Y_{i,k}) = 1 - \frac{2}{K}\sum_{k=1}^{K}\frac{\sum_{i=1}^{N_k} G_{i,k}Y_{i,k}}{\sum_{i=1}^{N_k} G_{i,k}^2 + \sum_{i=1}^{N_k} Y_{i,k}^2 + \epsilon}, \tag{7}$$

where K signifies the number of segmentation classes, N_k denotes the voxel count of class k, $G_{i,k}$ represents a binary value indicating whether class label k is the appropriate classification for pixel location i, and $Y_{i,k}$ corresponds to the probability of the associated prediction. ϵ is a very small positive number, referred to as the smoothing coefficient, which is configured to 10^{-7} in our experiments.

The issue of a significant discrepancy in the number of classes within an image sample presents an imbalanced classification problem. To address this, the WCE loss is utilized. WCE loss is calculated using a pixel-level softmax activation over the feature maps in tandem with cross-entropy. The predicted probability of class k at each pixel location $i \in \Omega$, where $\Omega \subset \mathbb{Z}^3$, is determined by inputting the predicted value into the softmax activation, which is defined as:

$$p_{i,k} = \frac{e^{Y_{i,k}}}{\sum_{k'=1}^{K} e^{Y_{i,k'}}}. \tag{8}$$

To effectively manage imbalanced classes, we determine the weight for each class present in the ground truth of the input image. This computation

[3]In certain literature [57, 58, 59], WCE is also referred to as weighted softmax loss.

is represented by the following equation:

$$w_k = 1 - \frac{\sum_{i=1}^{N} G_{i,k}}{\sum_{k'=1}^{K} \sum_{i=1}^{N} G_{i,k'}}, \tag{9}$$

where N refers to the total count of voxels in the ground truth, $\sum_{i=1}^{N} G_{i,k}$ signifies the count of pixels pertaining to class k in the ground truth, and $\sum_{k'=1}^{K} \sum_{i=1}^{N} G_{i,k'}$ stands for the aggregate number of pixels encompassing all classes in the ground truth. Subsequently, the WCE loss can be expressed as follows:

$$L_{WCE}(G_{i,k}, Y_{i,k}) = -\frac{1}{KN} \sum_{k=1}^{K} \sum_{i=1}^{N} w_k G_{i,k} \log(p_{i,k}). \tag{10}$$

Concurrently, WCE loss dynamically adjusts the weightage of different components in the image, reducing the significance of the background and bolstering the weight of specific internal classifications. This dynamic balancing aids in mitigating the effect of edge voxels on the overall loss. Consequently, both the $L_{decoder}$ loss, associated with the decoder's performance, and the L_{seg} loss, tied to the final segmentation output, are calculated using both Dice and WCE metrics with a weighted summation based on the λ value. They are defined as follows:

$$\begin{aligned} L_{decoder} =& \sum_{s=1}^{S} (L_{Dice}(G, P_s) \cdot \lambda + L_{Dice}(\bar{G}, P_s) \cdot (1 - \lambda)) + \\ & \sum_{s=1}^{S} (L_{WCE}(G, P_s) \cdot \lambda + L_{WCE}(\bar{G}, P_s) \cdot (1 - \lambda)), \\ L_{seg} =& L_{Dice}(G, Y_{seg}) \cdot \lambda + L_{Dice}(\bar{G}, Y_{seg}) \cdot (1 - \lambda) + \\ & L_{WCE}(G, Y_{seg}) \cdot \lambda + L_{WCE}(\bar{G}, Y_{seg}) \cdot (1 - \lambda), \end{aligned} \tag{11}$$

where P_s signifies the segmentation prediction generated from the low-level feature map at the sth stage of the decoder. G and $\bar{G}$ are the target labels of two samples in the new training sample pair after data augmentation by 3DMM-CutMix, respectively. The term S denotes the total number of stages in the decoder, which is set to 4 in our network design.

In conclusion, the overall loss function is formulated as follows:

$$L_{total} = a_d \times L_{decoder} + a_s \times L_{seg} + a_f \times L_{FDC} \tag{12}$$

where a_d, a_s, and a_f are the coefficients of the corresponding loss, respectively.

4. Experiments

4.1. Dataset

The first dataset used for method evaluation is the BraTS 2018 collection [61]. The dataset comprises 285 multi-contrast MRI scans, each containing four distinct MRI modalities: Fluid-Attenuated Inversion Recovery (FLAIR), T1-Contrast Enhanced (T1c), T1-Weighted (T1w), and T2-Weighted (T2w). Given that the annotations for the validation set are not publicly accessible, we partitioned the available 285 images into a training set of 190 images, with the remaining 95 used as a test set. The data split is the same as in [24, 33]. We normalized the modalities across all images by resizing each to a resolution of $128 \times 128 \times 128$. Our second dataset for method evaluation is the BraTS 2020 collection [62], which constitutes 369 multi-contrast MRI scans in four different MRI modalities. According to the same partitioning method as [24, 33], we divide the available 369 images into 246 training sets and 123 test sets according to a 2:1 ratio column, and the modalities of all images are similarly normalized to $128 \times 128 \times 128$ dimensions.

Each subject in these two datasets is represented through the above four MRI modalities, accompanied by voxel-level segmentation ground truth of three labels: *necrotic and non-enhancing tumor*, *edema*, and *enhancing tumor*. The training target is constructed by merging the three different tumor classes of the ground truth labels. Following the setting in previous work [63], we then converted each segmentation map into three binary maps, corresponding to three tumor categories: *Whole Tumor* (all three tumor classes), *Tumor Core* (all tumor classes excluding *edema*), and *Enhancing Tumor* (limited to the *enhancing tumor* class).

4.2. Implementation

We adopted the Dice Similarity Coefficient (DSC) as our evaluation metric [56]. Our proposed framework and compared methods were implemented with Python 3.8, and PyTorch 1.12 on one NVIDIA Tesla A100 GPU with one Intel Xeon Platinum 8358P CPU. Each input tensor representing a modality has dimensions $128 \times 128 \times 128 \times 1$, while the corresponding target tensor size is $128 \times 128 \times 128 \times 4$. For data augmentation, we followed previous work [33], applying random flipping, cropping, and intensity shifts. Subsequently, our proposed 3DMM-CutMix approach was also employed for model training. It's important to note that, during model testing, we didn't apply any data augmentation. We used the Adam optimizer with an initial learning

rate 0.0002 and a weight decay factor 0.0001. We maintain a batch size of 1 throughout our experiments. We set the total training period to 1000 epochs to ensure a fair comparison with other models. Our model is trained for about 78 hours with 80G memory on one GPU.

4.3. Parameter Analysis

The parameters involved are a_d, a_s, and a_f. We set a_d and a_s to 1.0 following [33]. To determine a suitable value for a_f, which balances the semantic segmentation, decoder, and distillation losses, we evaluated a range of values: 10, 1.0, 0.5, 0.1, and 0.05. The results in Table 1 demonstrate that smaller a_f values yielded higher segmentation performance. This indicates the semantic features may be more important for the Dice Similarity Coefficient metric. Based on these experiments, we set a_f to 0.1 by default to achieve a good trade-off between both metrics.

Table 1: The segmentation performance of our proposed IMS^{2}Trans using different values of the parameter a_f in Eq. (12), when all four modalities are used for training and testing with the 15 possible combinations of available modalities. The evaluation is conducted using Dice Similarity Coefficient (DSC) (%) to first find the results of each of the 15 different pattern combinations and then average them.

	Whole Tumor	Tumor Core	Enhancing Tumor
$a_f = 1 \times 10^{1}$	85.94	74.27	54.78
$a_f = 1 \times 10^{0}$	86.84	75.88	57.21
$a_f = 5 \times 10^{-1}$	**86.84**	75.89	58.01
$a_f = 1 \times 10^{-1}$	86.57	75.67	**58.28**
$a_f = 5 \times 10^{-2}$	86.60	**76.61**	57.83

4.4. Comparison of Segmentation Performance

For comparison, we selected six representative models that employ a single network, namely U-HeMIS [29], U-HVED [29], RobustSeg [28], ACN [24], D^2-Net [25], and mmFormer [33]. It is important to note that, for a fair comparison and consistency with the original papers [29, 24, 33], we use U-HeMIS in our experiments instead of HeMIS [64], which employs U-Net as the encoder of HeMIS. As RobustSeg [28] only conducted experiments on one BraTS 2015 collection, we obtained RobustSeg's BraTS 2018 results from the authors. Considering that the D^2-Net model was proposed after the ACN model, and both models use knowledge distillation to segment incomplete multi-modality brain tumors, we use the experimental results presented in previous paper [25] for both models on the BraTS 2018 collection. The

findings in Table 2 on the BraTS 2018 collection reveal that our model consistently outperforms U-HeMIS, U-HVED, RobustSeg, ACN, and D^2-Net across all three tumor classes in each of the 15 possible combinations of available modalities. On average, our model demonstrates superior segmentation performance compared to U-HeMIS, U-HVED, RobustSeg, ACN, D^2-Net, and mmFormer for all three tumor classes. These results highlight the effectiveness of our proposed model in handling missing modalities and achieving accurate segmentation in multi-modal MRI analysis. Then, we utilized ITK-SNAP [65], a software tool, to visualize the segmentation results for both the mmFormer model [33] and our proposed IMS^{2}Trans under 15 different missing modality conditions. The visualizations were presented in axial, sagittal, and coronal slice views, as depicted in Figure 4, where it clearly demonstrates that our IMS^{2}Trans outperforms mmFormer in the majority of the 15 combinations, showcasing a superior segmentation effect.

Table 2: Comparison of results in the BraTS 2018 collection achieved by our proposed IMS^{2}Trans and state-of-the-art unified models, including U-HeMIS, U-HVED, RobustSeg, ACN, D^2-Net, and mmFormer. The evaluation is conducted using Dice Similarity Coefficient (DSC) (%) across different combinations of modalities. In the caption, • and ∘ represent available and missing modalities, respectively, and 'F' denotes FLAIR in short.

Modalities				Whole Tumor							Tumor Core							Enhancing Tumor						
F	T1c	T1w	T2w	U-HeMIS	U-HVED	RobustSeg[28]	ACN[25]	D^2-Net[25]	mmFormer	Ours	U-HeMIS	U-HVED	RobustSeg[28]	ACN[25]	D^2-Net[25]	mmFormer	Ours	U-HeMIS	U-HVED	RobustSeg[28]	ACN[25]	D^2-Net[25]	mmFormer	Ours
•	∘	∘	∘	52.48	84.39	85.69	38.2	84.2	85.69	**85.81**	26.06	57.90	53.57	34.1	47.3	**64.45**	64.33	11.78	23.80	25.69	26.5	8.1	**34.58**	31.82
∘	•	∘	∘	61.53	53.62	73.31	41.3	42.8	**79.83**	79.52	65.29	59.59	76.83	35.8	65.1	**81.20**	80.72	62.02	57.64	67.07	28.9	66.3	72.31	**73.17**
∘	∘	•	∘	57.62	49.51	70.11	46.4	15.5	**78.55**	78.27	37.39	33.90	47.90	32.1	16.8	64.75	**64.93**	10.16	8.60	17.29	26.5	8.1	**34.82**	32.48
∘	∘	∘	•	80.96	79.83	82.24	28.9	76.3	85.10	**86.15**	57.20	54.67	57.49	30.7	56.7	**67.90**	67.52	25.63	22.82	28.97	34.2	16.0	**41.06**	36.42
•	•	∘	∘	68.99	85.93	**88.51**	50.9	87.5	87.88	88.09	71.49	75.07	80.62	45.1	80.8	80.93	**81.23**	66.10	68.36	70.30	36.5	64.8	71.99	**76.21**
•	∘	•	∘	64.62	85.71	**88.24**	58.3	87.3	87.96	88.03	41.12	61.14	60.68	45.8	61.6	70.20	**71.00**	10.71	27.96	32.13	36.6	9.5	39.73	**41.63**
•	∘	∘	•	82.95	87.58	88.28	44.7	87.9	**88.69**	88.14	57.68	62.70	61.16	47.2	62.6	**69.83**	69.70	30.22	32.31	33.84	**46.6**	17.4	41.49	41.83
∘	•	•	∘	68.47	64.22	77.18	58.6	62.1	**82.49**	82.33	72.46	67.55	78.72	47.7	78.2	81.22	**82.60**	66.22	61.11	69.06	38.8	70.7	75.53	**76.49**
∘	•	∘	•	82.48	81.32	85.19	49.8	84.1	87.80	**88.13**	76.64	73.92	80.20	51.2	80.3	**81.86**	81.75	67.83	67.83	69.71	47.4	68.7	72.58	**73.59**
∘	∘	•	•	82.41	81.56	84.78	55.3	80.1	87.20	**87.89**	60.92	56.26	62.19	47.3	63.2	**72.36**	70.90	32.39	24.29	32.01	**45.6**	16.5	45.56	40.99
•	•	•	∘	72.31	86.72	88.73	63.4	87.7	88.79	**89.02**	76.01	77.05	81.06	52.3	80.9	81.35	**82.55**	68.54	68.60	70.78	42.9	65.7	74.28	**76.03**
•	•	∘	•	83.85	88.09	89.27	56.3	88.8	**89.51**	89.47	77.53	76.75	80.72	56.6	80.7	80.98	**81.47**	68.72	68.93	70.88	53.1	66.4	73.03	**76.19**
•	∘	•	•	83.43	88.07	88.81	62.5	88.4	**89.29**	88.77	60.32	63.14	64.38	56.4	63.7	**72.45**	71.70	31.07	32.34	36.41	**52.5**	19.4	42.72	42.59
∘	•	•	•	83.94	82.32	86.01	64.8	80.9	88.26	**88.44**	78.96	75.28	80.33	59.0	79.0	81.85	**82.42**	69.92	67.75	70.10	53.8	68.3	74.26	**76.16**
•	•	•	•	84.74	88.46	89.45	67.6	88.8	89.83	**89.97**	79.48	77.71	80.86	61.7	80.1	81.36	**82.23**	70.24	69.03	71.13	56.6	68.4	73.00	**78.54**
Average				74.05	79.16	84.39	52.5	76.2	86.45	**86.57**	62.57	64.84	69.78	46.9	66.5	75.51	**75.67**	46.10	46.76	51.02	41.8	42.3	57.79	**58.28**

To quantitatively show the improvement, we selected 95 multimodal MRI images from the BraTS 2018 dataset as the test set. Experiments for brain tumor segmentation were then conducted on this test set and compared with the state-of-the-art mmFormer under 15 different conditions of missing modalities. For each method, we obtained segmentation results across all three tumor regions. Then, we leverage the Wilcoxon signed-rank test to show the statistical significance as it does not require the data to conform to any distribution. This yielded the p-values shown in Table 3, which indicate that for the Whole Tumor segmentation, 12 out of the 15 groups showed significant differences between mmFormer and IMS^{2}Trans ($p \leq 0.05$). For the Tumor Core, 3 out of 15 groups had $p \leq 0.05$. And for the Enhancing Tumor, 4 groups had $p \leq 0.05$. These results imply that IMS^{2}Trans can outperform

mmFormer at high statistical significance with much less computational cost.

Table 3: Statistical analysis of results in BraTS 2018 collection achieved by our proposed IMS^{2}Trans and state-of-the-art unified model mmFormer. The evaluation is conducted using p-value of Wilcoxon signed-rank test between mmFormer and IMS^{2}Trans across 15 different combinations of modalities. In the caption, • and ○ represent available and missing modalities, respectively, and 'F' denotes FLAIR in short.

Modalities				Whole Tumor	Tumor Core	Enhancing Tumor
F	T1c	T1w	T2w	p-value	p-value	p-value
●	○	○	○	0.0327	0.2250	0.9648
○	●	○	○	0.0267	0.4652	0.0999
○	○	●	○	0.6605	0.8266	0.9927
○	○	○	●	0.0014	0.4184	0.7248
●	●	○	○	0.0024	0.1347	0.0432
●	○	●	○	0.0038	0.6061	0.9818
●	○	○	●	0.0026	0.7054	0.5035
○	●	●	○	0.5874	0.0480	0.0441
○	●	○	●	0.0346	0.3287	0.3027
○	○	●	●	0.0086	0.7860	0.9875
●	●	●	○	0.0014	0.0397	0.1023
●	●	○	●	0.0004	0.1541	0.0711
●	○	●	●	0.0022	0.6793	0.9701
○	●	●	●	0.0834	0.0834	0.0247
●	●	●	●	0.0013	0.0193	0.0017

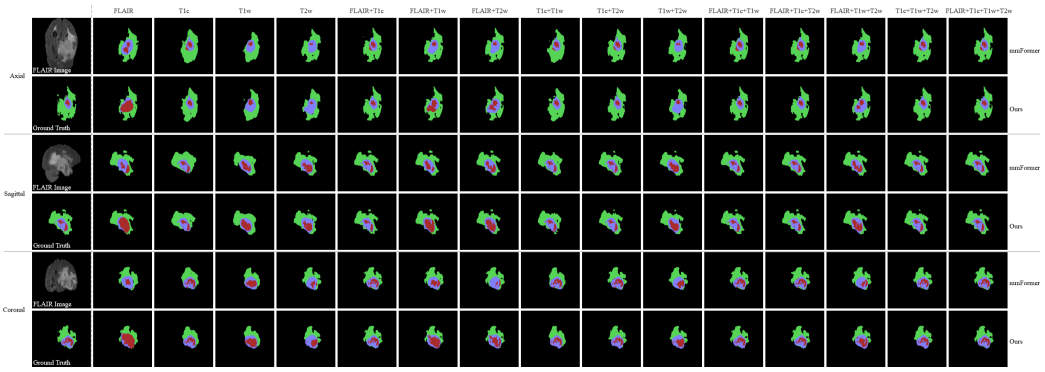


Figure 4: Visualization of segmentation results under 15 different missing modality conditions for mmFormer and our proposed IMS^{2}Trans compared to ground truth. The results are displayed in axial, sagittal, and coronal slice views. The colors represent different tumor classes: red for necrotic and non-enhancing tumor core, green for edema, and blue for enhancing tumor.

In order to further analyze the impact of the number of missing modalities on the segmentation performance of our model, we conducted experiments comparing our proposed IMS^{2}Trans with the mmFormer. Specifically, we focused on scenarios where only three modalities were available for training and testing, with the T1w modality being absent. The results of these experiments are presented in Table 4. It demonstrates that our IMS^{2}Trans model outperforms mmFormer in the segmentation of all three tumor classes across all seven possible combinations of available modalities, which is consistent with the scenarios of four modalities. Moreover, we provide visualizations of the segmentation results under the seven different missing modality conditions for both mmFormer and IMS^{2}Trans in axial, sagittal, and coronal slice views, as shown in Figure 5. These visualizations provide additional evidence of the superior segmentation performance of our IMS^{2}Trans model compared to mmFormer in the majority of the seven combinations.

Table 4: Results of our proposed IMS^{2}Trans and mmFormer models in BraTS 2018 collection when only three modalities are used for training and testing, with the T1w modality missing, where • and ○ represent available and missing modalities, respectively, and 'F' denotes FLAIR modality.

Modalities			Whole Tumor		Tumor Core		Enhancing Tumor	
F	T1c	T2w	mmFormer	Ours	mmFormer	Ours	mmFormer	Ours
•	○	○	74.63	**86.08**	43.07	**66.31**	18.04	**32.15**
○	•	○	60.53	**79.14**	56.68	**81.76**	53.94	**73.37**
○	○	•	66.03	**86.07**	44.57	**69.11**	18.47	**38.11**
•	•	○	81.80	**87.71**	66.61	**82.38**	61.04	**76.13**
•	○	•	82.41	**88.39**	54.78	**70.50**	23.08	**40.22**
○	•	•	77.01	**87.86**	66.98	**82.97**	59.50	**74.46**
•	•	•	84.55	**89.15**	69.97	**82.84**	62.51	**75.94**
Average			75.28	**86.34**	57.52	**76.56**	42.37	**58.62**

To better showcase the effectiveness of our method, we compared our IMS^{2}Trans to the current leading methods such as U-HeMIS, U-HVED, RobustSeg, RFNet [30], and mmFormer on the BraTS 2020 collection. Note that RobustSeg [28] has not performed experiments on the BraTS 2020 collection, so we use the experimental results from RobustSeg in [30]. Table 5 presents the experimental results, which indicate that our IMS^{2}Trans consistently performs better than U-HeMIS, U-HVED, RobustSeg, RFNet, and mmFormer across all three brain tumor categories, even with 15 possible missing modal combinations. These results demonstrate the effectiveness of our method in accurately segmenting brain tumors and handling missing

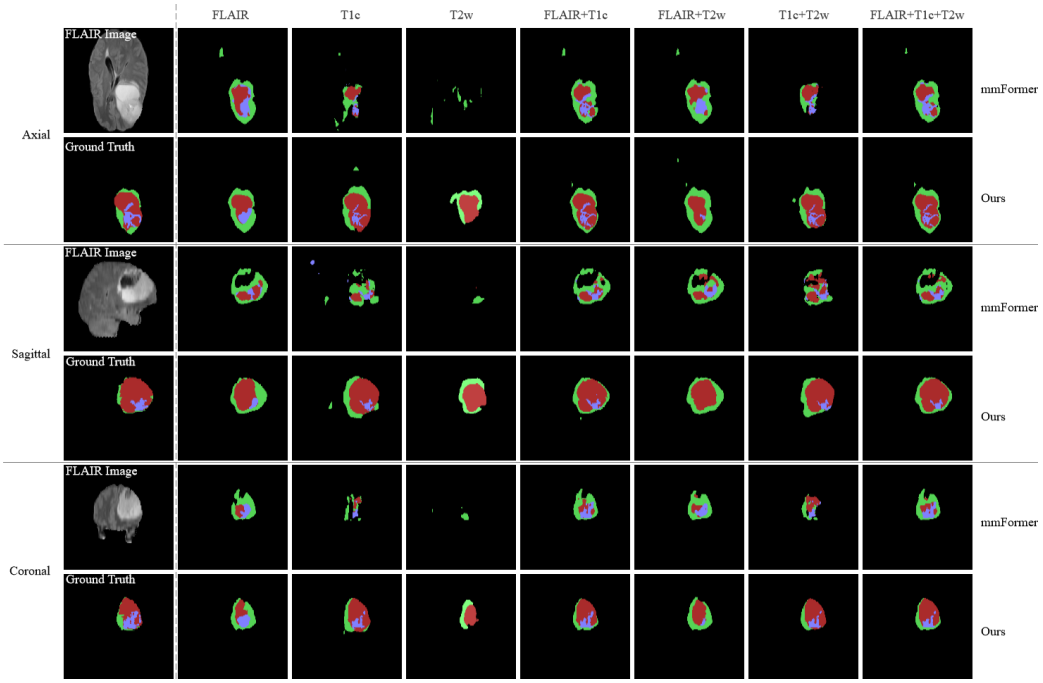


Figure 5: Visualization of the segmentation results under seven different missing modality conditions for mmFormer and our proposed IMS^{2}Trans compared to the ground truth. The results are shown in axial, sagittal, and coronal slice views. The color scheme represents different tumor classes: red for necrotic and non-enhancing tumor core, green for edema, and blue for enhancing tumor.

modalities.

Table 5: Comparison of results in BraTS 2020 collection achieved by our proposed IMS^{2}Trans and state-of-the-art unified models, including U-HeMIS, U-HVED, RobustSeg, RFNet and mmFormer. The evaluation is conducted using Dice Similarity Coefficient (DSC) (%) across different combinations of modalities. In the caption, • and ○ represent available and missing modalities, respectively, and 'F' denotes FLAIR in short.

Modalities				Whole Tumor						Tumor Core						Enhancing Tumor					
F	T1c	T1w	T2w	U-HeMIS	U-HVED	RobustSeg[30]	RFNet[30]	mmFormer	Ours	U-HeMIS	U-HVED	RobustSeg[30]	RFNet[30]	mmFormer	Ours	U-HeMIS	U-HVED	RobustSeg[30]	RFNet[30]	mmFormer	Ours
•	○	○	○	58.72	82.76	82.87	87.32	87.64	**88.47**	37.03	52.42	60.72	69.19	68.44	**69.72**	14.63	25.85	34.68	38.15	**42.20**	40.17
○	•	○	○	66.92	71.42	71.39	76.77	75.85	**79.31**	74.22	74.93	76.68	81.51	81.18	**82.76**	64.95	68.43	67.91	74.85	70.63	**75.60**
○	○	•	○	66.35	58.30	71.41	**77.16**	77.02	76.33	48.57	39.54	54.30	66.02	65.13	**66.74**	20.41	18.21	28.99	37.30	**39.04**	36.01
○	○	○	•	80.34	82.13	82.20	**86.05**	84.01	86.00	60.83	61.37	61.88	71.02	**71.05**	69.60	32.78	31.86	36.46	**46.29**	44.97	44.50
•	•	○	○	73.41	87.15	87.33	89.89	89.04	**90.75**	74.62	77.45	81.85	84.65	84.22	**84.96**	69.52	71.24	70.78	76.67	73.70	**77.39**
•	○	•	○	69.79	86.46	88.10	89.73	89.53	**90.23**	48.19	57.38	68.18	73.07	73.31	**76.08**	19.04	27.94	39.67	40.98	**45.64**	44.75
•	○	○	•	83.76	87.91	88.09	89.87	89.60	**90.45**	60.21	63.47	68.20	**74.14**	74.06	73.87	30.66	33.64	42.19	49.32	47.88	**49.42**
○	•	•	○	73.41	74.09	76.84	81.12	79.49	**81.17**	78.35	79.11	80.28	83.40	82.54	**84.39**	71.40	70.79	70.11	**78.01**	73.52	76.75
○	•	○	•	85.16	85.72	85.97	**87.74**	86.32	87.71	79.84	80.27	82.44	83.45	**84.71**	84.68	71.12	70.48	71.42	75.93	72.08	**77.48**
○	○	•	•	83.30	84.34	85.53	**87.73**	86.71	87.44	60.80	62.17	66.46	73.13	73.48	**73.95**	29.76	32.37	39.92	45.65	**47.92**	47.32
•	•	•	○	76.78	86.59	88.87	90.69	89.70	**90.91**	78.88	79.02	82.76	85.07	85.03	**86.12**	71.39	72.16	71.77	76.81	74.47	**78.17**
•	•	○	•	85.17	88.92	88.68	90.68	90.02	**91.17**	79.24	80.19	81.89	84.97	85.53	**85.54**	71.98	71.72	71.17	77.12	74.46	**77.72**
•	○	•	•	84.43	88.66	89.24	90.60	90.20	**90.80**	63.48	65.39	70.46	75.19	75.66	**76.39**	32.13	34.48	43.90	49.92	50.00	**50.02**
○	•	•	•	85.84	85.86	86.63	**88.25**	87.21	88.21	81.56	81.72	82.85	83.47	84.96	**85.57**	72.37	71.92	71.87	76.99	74.09	**78.16**
•	•	•	•	86.03	89.43	89.47	91.11	90.32	**91.13**	81.03	81.68	82.87	85.21	85.86	**86.04**	72.44	71.87	71.52	78.00	74.87	**78.17**
Average				77.29	82.65	84.17	86.98	86.17	**87.33**	67.12	69.07	73.45	78.23	78.34	**79.09**	49.64	51.53	55.49	61.47	60.36	**62.11**

4.5. Comparison of Model Complexity

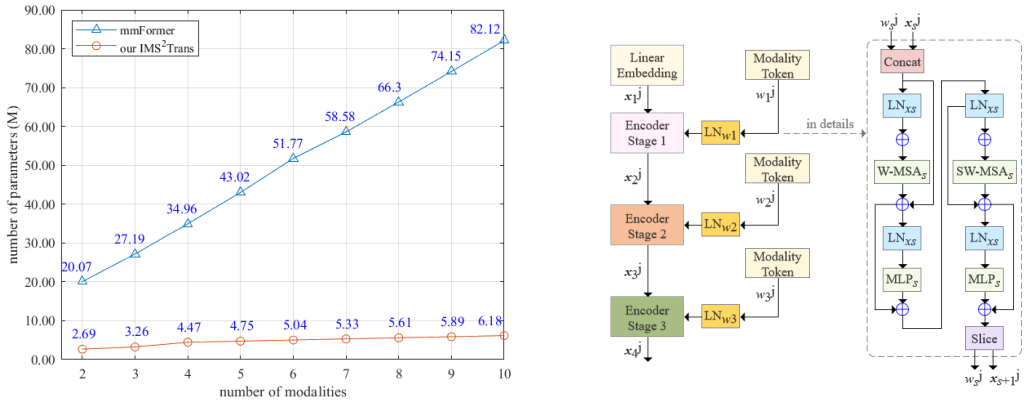
Furthermore, in order to provide a comprehensive analysis, we conducted an in-depth analysis to compare the efficiency of IMS^{2}Trans with the state-of-the-art mmFormer, as it achieved the best performance in Table 2. For a fair comparison, we specified the input size as $1\times4\times128\times128\times128$. The network configuration analysis is presented in Table 6, encompassing important metrics such as the number of parameters, FLOPs, model size, training speed per epoch, and inference speed per sample. The results demonstrate that our IMS^{2}Trans model exhibits significantly smaller computational and space complexity when compared to mmFormer. Specifically, our model comprises 4.47M parameters, 121.89G FLOPs, and a model size of 77.13MB. In contrast, mmFormer consists of 34.96M parameters, 265.79G FLOPs, and has a model size of 559.74MB. These metrics clearly indicate that our IMS^{2}Trans model achieves a notable reduction in both computational complexity and space complexity compared to mmFormer. The smaller computational and space complexity of IMS^{2}Trans provides several advantages. Firstly, it allows for more efficient model training, resulting in faster convergence during the training process. Secondly, it facilitates faster inference speed, enabling real-time or near-real-time applications in clinical settings. Lastly, the reduced model size leads to lower memory requirements, making our model more feasible for deployment on resource-constrained devices or in settings where high-performance computational resources may not be readily available.

We also conducted an analysis to highlight the effect of the number of missing modalities on space complexity. This is illustrated by comparing the number of parameters between the state-of-the-art mmFormer model and our IMS^{2}Trans in terms of handling the different number of modalities, as depicted in Figure 6(a). As shown in the figure, the number of parameters in the mmFormer model increases linearly as the number of available modalities increases from two to ten, ranging from 20M to 82M. In contrast, our IMS^{2}Trans exhibits a relatively stable number of parameters, varying from 2.69M to 6.18M regardless of the number of available modalities. With a fixed number of parameters, our model remains lightweight and avoids exponential growth in parameter count as the number of modalities increases. This makes our model more scalable and efficient in handling various missing modality scenarios. The relatively stable number of parameters in IMS^{2}Trans is particularly beneficial in real clinical scenarios where a varying number of modalities may be available due to equipment limitations or data acquisition challenges. By maintaining a consistent model size, our approach ensures

that the computational resources required for model training and inference remain manageable and practical.

Table 6: Efficiency comparisons of the proposed IMS^{2}Trans with mmFormer.

Network	Params (M)	FLOPs (G)	Model Size (MB)	Train. Speed (s)	Inf. Speed (ms)
mmFormer	34.96	265.79	559.74	288.91	1892
Ours	4.47	121.89	77.13	269.80	1649



(a) Number of parameters when different number of modalities involved

(b) Concatenate design of modality token

Figure 6: (a) Comparison of the number of parameters for mmFormer and our proposed IMS^{2}Trans models in terms of different number of involved modalities. (b) Illustration of the concatenate design in the swin transformer block with modality token.

4.6. Ablation Study

In order to evaluate the design of modality token, we conducted a comparison between two different strategies in terms of training parameters, GPU performance, and segmentation performance in Table 7. The first design approach, which we adopted in our network, is the additive implementation, as illustrated in Figure 3(a). An alternative approach is the concatenate implementation as depicted in Figure 6(b). In the concatenate implementation, the feature token x_s^j of the j-th modality and the corresponding modality token w_s^j are first concatenated to form a new token. This new token is then passed through two consecutive window-based multi-head self-attention modules (W-MSA and SW-MSA) within the swin transformer block. Finally, a

slicing module is used to extract the updated feature token x_{s+1}^j, discarding the modality token w_s^j in the output. Unlike the additive operation, the concatenate operation does not modify the internal modules of the swin transformer block. Instead, it introduces a new token at the input and extracts a modality token at the output for subsequent removal. In Table 7, the results yield that the additive implementation outperforms the concatenate implementation in all aspects, including training parameters, GPU performance, and average segmentation performance. Therefore, our IMS^{2}Trans network adopts the additive implementation due to its superior performance. By choosing the additive implementation, we ensure that the swin transformer block with modality token effectively captures the interdependencies among modalities and produces more accurate and reliable segmentation results.

Table 7: Comparison of two design approaches for the modality token, namely the additive and concatenation schemes.

Additive	Concatenate	Params (M)	FLOPs (G)	Model Size (MB)	Whole Tumor	Tumor Core	Enhancing Tumor
✓		4.47	121.89	77.13	86.57	75.67	58.28
	✓	17.12	147.70	279.51	85.85	75.31	58.05

To demonstrate the effectiveness of our proposed 3DMM-CutMix method, we provide visualizations of the data augmentation results in Figure 7. The figure consists of four rows, each representing FLAIR, T1c, T1w, and T2w images, respectively. Within each row, there are six columns depicting different stages of the data augmentation process. In the first column, the source image is displayed, while the second column shows the corresponding ground truth for the source image. The third column represents the reference image, and the fourth column displays the ground truth for the reference image. The fifth column showcases the 3D image cropped from the reference image using the 3DMM-CutMix method. This cropped image serves as a mixing component in the data augmentation process. Finally, in the sixth column, we present the resulting image obtained by blending the source image from the first column with the cropped reference image from the fifth column using the 3DMM-CutMix method. These visualizations provide an intuitive representation of how our 3DMM-CutMix method effectively combines information from the source and reference images to enhance the training process. By generating augmented images that incorporate relevant features from both the source and reference images, the 3DMM-CutMix method promotes better generalization and improves the robustness of the deep learning model for MRI analysis. Overall, the visualization results in Figure 7 demonstrate

the effectiveness and potential of our proposed 3DMM-CutMix method as a valuable data augmentation technique in deep learning-based MRI analysis.

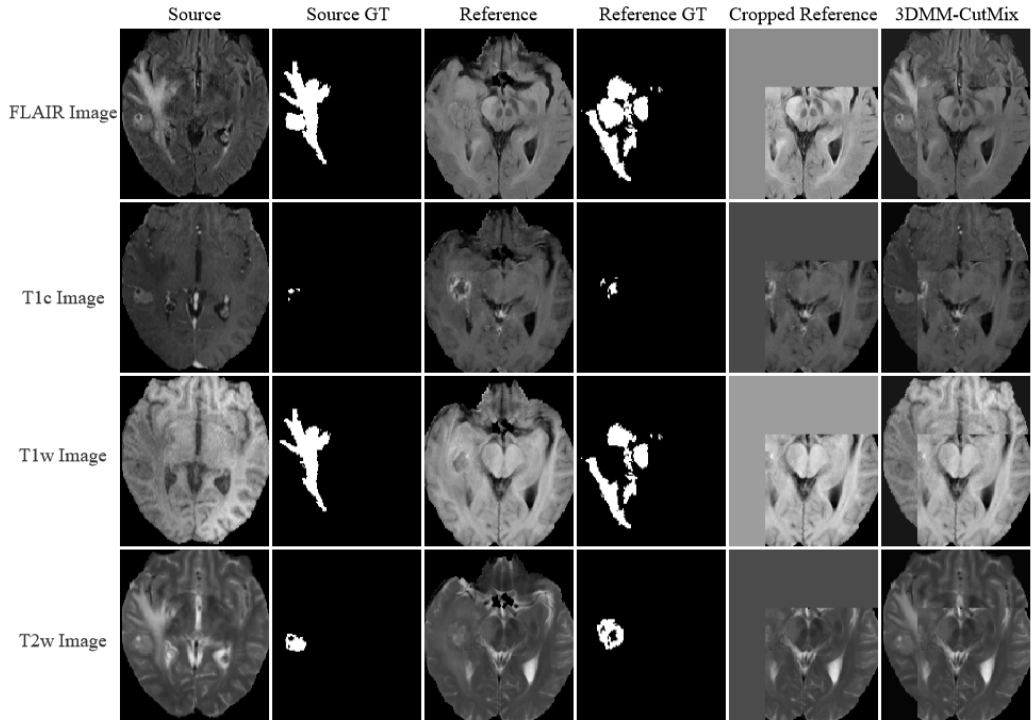


Figure 7: Visualization of the data augmentation results using our proposed 3DMM-CutMix method. 'GT' denotes ground truth.

To gain a better understanding of the contributions of different components in our IMS^{2}Tran network, we conducted an ablation study. Specifically, we evaluated the impact of 3DMM-CutMix, swin transformer with modality token, feature distillation loss, shifted MLP bottleneck, and decoder loss on training parameters, GPU performance, and average segmentation performance. The results of this study are summarized in Table 8, where the first column corresponds to the 3DMM-CutMix functional module. When this module is removed, it indicates that there is only one target label for each sample, which can be achieved by setting λ to 1 in Eq. (11). The second column represents the presence or absence of the swin transformer with modality token. If this component is removed, the model solely relies on the swin transformer without modality tokens. The third column indicates the inclusion or exclusion of the feature distillation module. When removed, it implies the

absence of the corresponding feature distillation loss, L_{FDC}, in the model. The fourth column signifies the removal of both the intra-modal shifted MLP on the left and the inter-modal shifted MLP on the right in Figure 2. Lastly, the fifth column pertains to the decoder loss. If this loss is removed, it solely affects the calculation of the total loss. From the results presented in Table 8, it is evident that the absence of the 3DMM-CutMix module leads to a drastic decrease in segmentation performance for all three tumor regions. Additionally, the inclusion of the swin transformer with modality token, feature distillation loss, shifted MLP bottleneck, and decoder loss contribute to performance improvements across the whole tumor region, tumor core region, and enhancing tumor region. These findings demonstrate the crucial roles played by the various components in our IMS^{2}Tran network. The 3DMM-CutMix module significantly enhances the model's ability to handle missing modalities, while the swin transformer with modality token, feature distillation loss, shifted MLP bottleneck, and decoder loss collectively contribute to improved segmentation performance. In addition, Table 8 also shows that except for the reduction of model parameters and GPU performance caused by the swin transformer without modality token, the lack of other modules has little impact. The reduction of model parameters is not large because the shifted MLP bottleneck is a lightweight module. The 3DMM-CutMix and feature distillation modules do not involve learnable parameters, so they have no effect on the number of model parameters. The lack of decoder loss can only reduce some convolutional layers and upsampling layers, so it also has little impact on the number of model parameters. Overall, these ablation study results validate the effectiveness and importance of each component in our proposed network architecture.

Table 8: Ablation study. ✓ indicates that the module is included in our model, and ✗ indicates that it is not included. '3D-C' denotes 3DMM-CutMix, 'MT' denotes modality token, and 'FD' denotes feature distillation. 'S-MLP' denotes shifted MLP bottleneck, '$L_{decoder}$' denotes the loss associated with the decoder's performance.

3D-C	MT	FD	S-MLP	$L_{decoder}$	Params (M)	FLOPs (G)	Model Size (MB)	Whole Tumor	Tumor Core	Enhancing Tumor
✓	✓	✓	✓	✓	4.47	121.89	77.13	86.57	75.67	58.28
✗	✓	✓	✓	✓	4.47	121.89	77.13	74.04(↓)	62.88(↓)	47.31(↓)
✓	✗	✓	✓	✓	3.33	121.62	63.42	86.45(↓)	75.86(↑)	56.18(↓)
✓	✓	✗	✓	✓	4.47	121.89	77.13	85.74(↓)	75.18(↓)	57.19(↓)
✓	✓	✓	✗	✓	4.46	121.76	77.04	86.20(↓)	76.45(↑)	57.08(↓)
✓	✓	✓	✓	✗	4.47	121.85	77.10	86.09(↓)	76.43(↑)	56.79(↓)

4.7. Discussion

Tables 2, 4, and 5 evidently yield that the best-performing existing approaches in brain tumor segmentation, although tailored for the incomplete modalities, still assume the availability of all modal data during training and subsequently simulate scenarios with missing modalities to create a unified model. However, when faced with the circumstance of incomplete modalities, these methods still demonstrate poor performance. In contrast, our IMS^{2}Tran model is better suited for situations where not all modalities are available, outperforming state-of-the-art. This superiority can be attributed to several key factors. First and foremost, our model benefits from an effective data augmentation strategy 3DMM-CutMix, which plays a crucial role in guiding the model to focus on both local and global objectives. Another significant contribution of our model is the adoption of a shared parameter encoder, allowing our model to learn common features from different modalities. At the bottleneck stage of IMS^{2}Tran, we leverage the complementary information present among the available modalities. This enables us to obtain a unified representation in the latent space, which proves instrumental in enhancing the overall representation capability of the feature map. Through feature distillation methods, we refine the learned features by leveraging the interdependencies among the modalities, resulting in a more comprehensive and discriminative representation. For the decoder stage, our model not only delivers accurate segmentation results but also facilitates supervised learning between each available modality and the target. This is accomplished through the implementation of skip connections, allowing information to flow not only within the model but also between different modalities. By establishing these connections, our model effectively enhances its adaptability to single-modal scenarios, further bolstering its robustness and generalization capabilities.

Another notable observation is that existing state-of-the-art models have been increasing in computational and space complexity in order to improve the performance of brain tumor segmentation. This is primarily reflected in the growing number of parameters and size indicators of these models, as shown in Table 6 and Figure 6 (a). However, IMS^{2}Tran addresses this issue and significantly reduces space complexity by a parameter-shared encoder and lightweight shifted MLP bottleneck. Consequently, compared to the state-of-the-art mmFormer, which has a linear complexity to the number of modalities, the complexity of our IMS^{2}Tran to the number of modalities remains in a fixed range.

Briefly, the key advantage of IMS^{2}Tran lies in the ability to maintain the performance of brain tumor segmentation for MRI images while significantly reducing the model's overall complexity. IMS^{2}Tran demonstrated its promising future for real clinical scenarios where incomplete modalities are common and high-performance GPU devices may not be readily available.

5. Conclusion

It is common to encounter incomplete modalities of MRI scans in clinical practice, leading to a decline in the performance of deep learning driven solutions. Hence, in this work, we propose IMS^{2}Trans, an innovative lightweight, and scalable network architecture specifically designed for brain tumor segmentation from incomplete modalities. IMS^{2}Trans takes a unique stance by extracting latent feature maps from all observed modalities using a single shared encoder. Our method offers significant advantages in terms of scalability compared to existing methods that rely on multiple encoders, where the number of encoders equals the number of involved modalities. The use of a single encoder not only reduces computational complexity but also simplifies the overall model structure, making it easier to manage and optimize. Furthermore, the unified extraction of latent feature maps through a single encoder facilitates better information sharing and fusion among the modalities, leading to improved segmentation performance even when modalities are missing.

The future direction will encompass three main aspects to further advance the field. Firstly, the segmentation performance can be enhanced by exploring semi-supervised or unsupervised learning approaches, by leveraging unlabeled data or additional auxiliary tasks to improve the model's ability to handle missing modalities effectively. Secondly, the computational complexity can be further alleviated by incorporating efficient operators including FlashAttention [66] to replace the vanilla self-attention heads. Lastly, we plan to generalize the application of our model beyond MRI segmentation tasks. The modular nature and flexibility of IMS^{2}Trans make it highly adaptable to other medical tasks such as anomaly detection and disease diagnosis in other imaging modalities, such as CT, and X-rays.

Declaration of Competing Interest

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

paper.

Acknowledgments

This work is supported by the Henan Key Research Projects of Higher Education Institutions of China (No. 24A520041).

References

[1] I. Mecheter, M. Abbod, A. Amira, H. Zaidi, Deep learning with multiresolution handcrafted features for brain mri segmentation, Artificial intelligence in medicine 131 (2022) 102365.

[2] A. Delli Pizzi, A. M. Chiarelli, P. Chiacchiaretta, M. d'Annibale, P. Croce, C. Rosa, D. Mastrodicasa, S. Trebeschi, D. M. J. Lambregts, D. Caposiena, et al., Mri-based clinical-radiomics model predicts tumor response before treatment in locally advanced rectal cancer, Scientific Reports 11 (1) (2021) 1–11.

[3] W. Yao, C. Liu, N. Wang, H. Zhou, H. Chen, W. Qiao, Anisamide-modified dual-responsive drug delivery system with mri capacity for cancer targeting therapy, Journal of Molecular Liquids 340 (2021) 116889.

[4] W. Yao, C. Liu, N. Wang, H. Zhou, H. Chen, W. Qiao, An mri-guided targeting dual-responsive drug delivery system for liver cancer therapy, Journal of Colloid and Interface Science 603 (2021) 783–798.

[5] R. Azad, N. Khosravi, M. Dehghanmanshadi, J. Cohen-Adad, D. Merhof, Medical image segmentation on mri images with missing modalities: A review, arXiv preprint arXiv:2203.06217 (2022).

[6] M. H. Hesamian, W. Jia, X. He, P. Kennedy, Deep learning techniques for medical image segmentation: achievements and challenges, Journal of digital imaging 32 (2019) 582–596.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[8] Y. Li, Z. Wang, L. Yin, Z. Zhu, G. Qi, Y. Liu, X-net: a dual encoding–decoding method in medical image segmentation, The Visual Computer 39 (2023) 2223–2233.

[9] Y. Xu, X. He, G. Xu, G. Qi, K. Yu, L. Yin, P. Yang, Y. Yin, H. Chen, A medical image segmentation method based on multi-dimensional statistical features, Frontiers in Neuroscience 16 (2022) 1009581.

[10] Z. Zhu, X. He, G. Qi, Y. Li, B. Cong, Y. Liu, Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal mri, Information Fusion 91 (2023) 376–387.

[11] X. He, G. Qi, Z. Zhu, Y. Li, B. Cong, L. Bai, Medical image segmentation method based on multi-feature interaction and fusion over cloud computing, Simulation Modelling Practice and Theory 126 (2023) 102769.

[12] Y. Lu, Y. Chang, Z. Zheng, Y. Sun, M. Zhao, B. Yu, C. Tian, Y. Zhang, Gmetanet: Multi-scale ghost convolutional neural network with auxiliary metaformer decoding path for brain tumor segmentation, Biomedical Signal Processing and Control 83 (2023) 104694.

[13] H. Liu, G. Huo, Q. Li, X. Guan, M.-L. Tseng, Multiscale lightweight 3d segmentation algorithm with attention mechanism: Brain tumor image segmentation, Expert Systems with Applications 214 (2023) 119166.

[14] M. J. Graves, D. G. Mitchell, Body mri artifacts in clinical practice: a physicist's and radiologist's perspective, Journal of Magnetic Resonance Imaging 38 (2) (2013) 269–287.

[15] B. M. Dale, M. A. Brown, R. C. Semelka, MRI: basic principles and applications, John Wiley & Sons, 2015.

[16] K. G. Hollingsworth, Reducing acquisition time in clinical mri by data undersampling and compressed sensing reconstruction, Physics in Medicine & Biology 60 (21) (2015) R297.

[17] A. S. Chaudhari, C. M. Sandino, E. K. Cole, D. B. Larson, G. E. Gold, S. S. Vasanawala, M. P. Lungren, B. A. Hargreaves, C. P. Langlotz,

Prospective deployment of deep learning in mri: a framework for important considerations, challenges, and recommendations for best practices, Journal of Magnetic Resonance Imaging 54 (2) (2021) 357–371.

[18] L. Zimmermann, B. Knäusl, M. Stock, C. Lütgendorf-Caucig, D. Georg, P. Kuess, An mri sequence independent convolutional neural network for synthetic head ct generation in proton therapy, Zeitschrift für Medizinische Physik 32 (2) (2022) 218–227.

[19] T. Zhou, S. Ruan, H. Hu, A literature survey of mr-based brain tumor segmentation with missing modalities, Computerized Medical Imaging and Graphics (2022) 102167.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.

[21] L. Jiang, Y. Mao, X. Chen, X. Wang, C. Li, Cola-diff: Conditional latent diffusion model for multi-modal mri synthesis, arXiv preprint arXiv:2303.14081 (2023).

[22] Q. Wang, L. Zhan, P. Thompson, J. Zhou, Multimodal learning with incomplete modalities by knowledge distillation, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1828–1838.

[23] S. Vadacchino, R. Mehta, N. M. Sepahvand, B. Nichyporuk, J. J. Clark, T. Arbel, Had-net: A hierarchical adversarial knowledge distillation network for improved enhanced tumour segmentation without post-contrast images, in: Medical Imaging with Deep Learning, PMLR, 2021, pp. 787–801.

[24] Y. Wang, Y. Zhang, Y. Liu, Z. Lin, J. Tian, C. Zhong, Z. Shi, J. Fan, Z. He, Acn: adversarial co-training network for brain tumor segmentation with missing modalities, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24, Springer, 2021, pp. 410–420.

[25] Q. Yang, X. Guo, Z. Chen, P. Y. Woo, Y. Yuan, D2-net: Dual disentanglement network for brain tumor segmentation with missing modalities, IEEE Transactions on Medical Imaging 41 (10) (2022) 2953–2964.

[26] R. Azad, N. Khosravi, D. Merhof, Smu-net: Style matching u-net for brain tumor segmentation with missing modalities, in: International Conference on Medical Imaging with Deep Learning, PMLR, 2022, pp. 48–62.

[27] M. Havaei, N. Guizard, N. Chapados, Y. Bengio, Hemis: Hetero-modal image segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, Springer, Cham, 2016, pp. 469–477.

[28] 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: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22, Springer, 2019, pp. 447–456.

[29] R. Dorent, S. Joutard, M. Modat, S. Ourselin, T. Vercauteren, Hetero-modal variational encoder-decoder for joint modality completion and segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22, Springer, 2019, pp. 74–82.

[30] Y. Ding, X. Yu, Y. Yang, Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 3975–3984.

[31] 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.

[32] Y. Shen, Personalized stain style transfer layers for distributed histology classification, in: Medical Imaging 2022: Digital and Computational Pathology, Vol. 12039, SPIE, 2022, pp. 134–139.

[33] 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: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Springer Nature Switzerland, Cham, 2022, pp. 107–117.

[34] T. Zhou, Feature fusion and latent feature learning guided brain tumor segmentation and missing modality recovery network, Pattern Recognition 141 (2023) 109665.

[35] A. Konwer, X. Hu, J. Bae, X. Xu, C. Chen, P. Prasanna, Enhancing modality-agnostic representations via meta-learning for brain tumor segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 21415–21425.

[36] 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: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.

[37] J. M. J. Valanarasu, V. M. Patel, Unext: Mlp-based rapid medical image segmentation network, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V, Springer, 2022, pp. 23–33.

[38] S. Atito, M. Awais, J. Kittler, Sit: Self-supervised vision transformer, arXiv preprint arXiv:2104.03602 (2021).

[39] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: Regularization strategy to train strong classifiers with localizable features, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6023–6032.

[40] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, T. Cukur, Image synthesis in multi-contrast mri with conditional generative adversarial networks, IEEE transactions on medical imaging 38 (10) (2019) 2375–2388.

[41] B. Yu, L. Zhou, L. Wang, Y. Shi, J. Fripp, P. Bourgeat, Ea-gans: edge-aware generative adversarial networks for cross-modality mr image synthesis, IEEE transactions on medical imaging 38 (7) (2019) 1750–1762.

[42] M. Yurt, S. U. Dar, A. Erdem, E. Erdem, K. K. Oguz, T. Çukur, mustgan: multi-stream generative adversarial networks for mr image synthesis, Medical image analysis 70 (2021) 101944.

[43] A. Sharma, G. Hamarneh, Missing mri pulse sequence synthesis using multi-modal generative adversarial network, IEEE Transactions on Medical Imaging 39 (4) (2020) 1170–1183.

[44] Y. Zhang, C. Peng, Q. Wang, D. Song, K. Li, S. K. Zhou, Unified multi-modal image synthesis for missing modality imputation (2023). arXiv:2304.05340.

[45] H. Yang, J. Sun, Z. Xu, Learning unified hyper-network for multi-modal mr image synthesis and tumor segmentation with missing modalities, IEEE Transactions on Medical Imaging (2023).

[46] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784 (2014).

[47] F.-A. Croitoru, V. Hondru, R. T. Ionescu, M. Shah, Diffusion models in vision: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).

[48] Y. Shen, L. Xu, Y. Yang, Y. Li, Y. Guo, Mixed sample augmentation for online distillation, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[49] Y. Shen, Y. Zhou, L. Yu, Cd2-pfed: Cyclic distillation-guided channel decoupling for model personalization in federated learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10041–10050.

[50] Y. Shen, L. Xu, Y. Yang, Y. Li, Y. Guo, Self-distillation from the last mini-batch for consistency regularization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11943–11952.

[51] Y. Shen, L. Xu, Y. Yang, Y. Li, Y. Guo, Online distillation with mixed sample augmentation, arXiv preprint arXiv:2206.12370 (2022).

[52] G. Huang, Y. Sun, Z. Liu, D. Sedra, K. Q. Weinberger, Deep networks with stochastic depth, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, Springer, 2016, pp. 646–661.

[53] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

[54] D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus), arXiv preprint arXiv:1606.08415 (2016).

[55] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR, 2020, pp. 1597–1607.

[56] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 fourth international conference on 3D vision (3DV), Ieee, 2016, pp. 565–571.

[57] H. Zhang, X. Wang, Z. He, Weighted softmax loss for face recognition via cosine distance, in: Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13, Springer, 2018, pp. 340–348.

[58] F. Jia, Y. Lei, N. Lu, S. Xing, Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization, Mechanical Systems and Signal Processing 110 (2018) 349–367.

[59] J. Wang, C. Zheng, X. Yang, L. Yang, Enhanceface: Adaptive weighted softmax loss for deep face recognition, IEEE Signal Processing Letters 29 (2021) 65–69.

[60] 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.

[61] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, et al., Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge, arXiv preprint arXiv:1811.02629 (2018).

[62] 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.

[63] L. Weninger, O. Rippel, S. Koppers, D. Merhof, Segmentation of brain tumors and patient survival prediction: methods for the brats 2018 challenge, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4, Springer, 2019, pp. 3–12.

[64] M. Havaei, N. Guizard, N. Chapados, Y. Bengio, Hemis: Hetero-modal image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19, Springer, 2016, pp. 469–477.

[65] P. A. Yushkevich, Y. Gao, G. Gerig, Itk-snap: An interactive tool for semi-automatic segmentation of multi-modality biomedical images, in: 2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC), IEEE, 2016, pp. 3342–3345.

[66] T. Dao, D. Fu, S. Ermon, A. Rudra, C. Ré, Flashattention: Fast and memory-efficient exact attention with io-awareness, Advances in Neural Information Processing Systems 35 (2022) 16344–16359.