IM-Fuse: A Mamba-based Fusion Block for Brain Tumor Segmentation with Incomplete Modalities

Vittorio Pipoli[⋆,1,2], Alessia Saporita[⋆,1,3], Kevin Marchesini[⋆,1],
Costantino Grana[1], Elisa Ficarra[†,1], and Federico Bolelli[†,1]🖂

[1] University of Modena and Reggio Emilia, Italy
[2] University of Pisa, Italy
[3] University of Bologna, Italy

Abstract. Brain tumor segmentation is a crucial task in medical imaging that involves the integrated modeling of four distinct imaging modalities to identify tumor regions accurately. Unfortunately, in real-life scenarios, the full availability of such four modalities is often violated due to scanning cost, time, and patient condition. Consequently, several deep learning models have been developed to address the challenge of brain tumor segmentation under conditions of missing imaging modalities. However, the majority of these models have been evaluated using the 2018 version of the BraTS dataset, which comprises only 285 volumes. In this study, we reproduce and extensively analyze the most relevant models using BraTS2023, which includes $1,251$ volumes, thereby providing a more comprehensive and reliable comparison of their performance. Furthermore, we propose and evaluate the adoption of Mamba as an alternative fusion mechanism for brain tumor segmentation in the presence of missing modalities. Experimental results demonstrate that Transformer-based architectures achieve leading performance on BraTS2023, outperforming purely convolutional models that were instead superior in BraTS2018. Meanwhile, the proposed Mamba-based architecture exhibits promising performance in comparison to state-of-the-art models, competing and even outperforming Transformers. The source code is publicly released alongside the benchmark developed for the evaluation: `https://github.com/AImageLab-zip/IM-Fuse`.

Keywords: Brain Tumor Segmentation · Missing Modalities · Mamba

1 Introduction

In recent years, deep learning methods have significantly advanced the state-of-the-art in medical image segmentation across various imaging modalities [8, 23]. While some tasks involve well-defined anatomical structures and can be accurately performed using a single imaging modality [4,5,21,22], others require multiple imaging sources to account for lesion heterogeneity, enhance contrast between sub-regions, and capture complex anatomical variations [15,33,37].

[⋆] Equal contribution. Authors are allowed to list their name first on their CVs.
[†] Equal supervision. 🖂 Corresponding author: `federico.bolelli@unimore.it`

The current gold standard for clinical imaging diagnosis of brain tumors is multi-parametric Magnetic Resonance Imaging (MRI) [14], which is critical for accurate delineation and volume quantification, therapy planning, and follow-up [3]. Usually, four modalities providing complementary information and supporting tumor sub-region analysis are employed: FLuid-Attenuated Inversion Recovery (FLAIR), T1-weighted images (T1), T1-weighted images with contrast enhancement (T1c), and T2-weighted images (T2). In clinical practice, specific modalities can be absent due to different acquisition protocols, image corruption, scanner availability, or patient conditions, e.g., allergies to certain contrast materials. Consequently, recent research has focused on methods for compensating for missing modalities [40], a challenge encountered not only in medical imaging but also in broader visual-recognition tasks [18, 24]. In most of the existing MRI-related literature, each modality is processed by a dedicated encoder, and a shared decoder uses the extracted features, usually fused or aligned in the bottleneck, for the final segmentation. These methods can be categorized into four groups: reconstruction-based approaches, latent-space feature fusion, knowledge distillation, and domain adaptation.

Reconstruction-based approaches aim to compensate for missing MRI sequences by synthesizing them in the original volume space, like M^3AE [19], or in the latent-feature space, such as M^3FeCon [36].

Latent-space features fusion methods address the challenge of missing modalities by learning a shared representation in the latent space, which remains robust even when certain modalities are absent. First works explored simple methods for the fusion which use statistics of the modalities [13] and variational autoencoder, e.g., U-HVED [10] and DRM-VAE [41], then more sophisticated works came out leveraging on Transformers, such as mmFormer [38] and MFTrans [28], other specific types of attention mechanisms [9, 39], or auxiliary tasks with a combination of different losses like ShaSpec [31]. Latent-space features fusion is sometimes used in combination with reconstruction, the latter as a secondary task to enhance the encoders feature extraction [6, 20].

Knowledge distillation-based methods employ a teacher-student approach, where a teacher network trained on full-modality data guides one or more student networks with incomplete inputs [1, 7, 16, 26, 29, 30, 32, 35]. While these methods often yield high accuracy by transferring rich knowledge from teacher to student, they are sensitive to teacher quality and require additional training overhead.

Lastly, *domain adaptation-based* approaches attempt to mitigate distribution shifts caused by missing modalities through adversarial or alignment strategies, either as standalone methods [17, 27] or in combination with other paradigms, such as ACN [34], which integrates adversarial learning with knowledge distillation, and the work by Qiu *et al.* [25], which employs latent feature fusion and prompting within a domain adaptation framework.

Most of the aforementioned methods have been developed and tested on BraTS2018 [3], provided by the homonymous MICCAI challenge, and comprising a total of 285 volumes. However, in the context of brain tumor segmentation, Transformer-based solutions struggle to compete with classical convolutional models, mainly because of data scarcity. A significant part of our effort is devoted to building a new benchmark based on the newest BraTS2023[4] [2], which includes $1,251$ volumes, thereby providing a more comprehensive and reliable comparison of state-of-the-art existing methods. Our experiments demonstrate that Transformer-based architectures take advantage of the increased number of training samples of BraTS2023, overtaking pure convolutional models.

Besides the extensive evaluation of state-of-the-art solutions on BraTS2023, we propose IM-Fuse (Incomplete Modality Fusion), a novel segmentation framework based on latent-space feature fusion leveraging the Mamba architecture [12], a recently developed state space model that has significantly influenced both natural language processing and computer vision fields. Notably, Mamba is characterized by its high memory retention capabilities, which facilitate exceptional long-context reasoning, and by a selective mechanism that enables the propagation or discarding of input tokens based on their semantic content. Accordingly, in this paper, we propose a novel Interleaved Mamba Fusion Block (I-MFB) that harnesses Mamba's capabilities to integrate multimodal information and handle sparse inputs resulting from missing modalities (Fig. 1), an issue affecting most of the existing architectures, thereby outperforming state-of-the-art models.

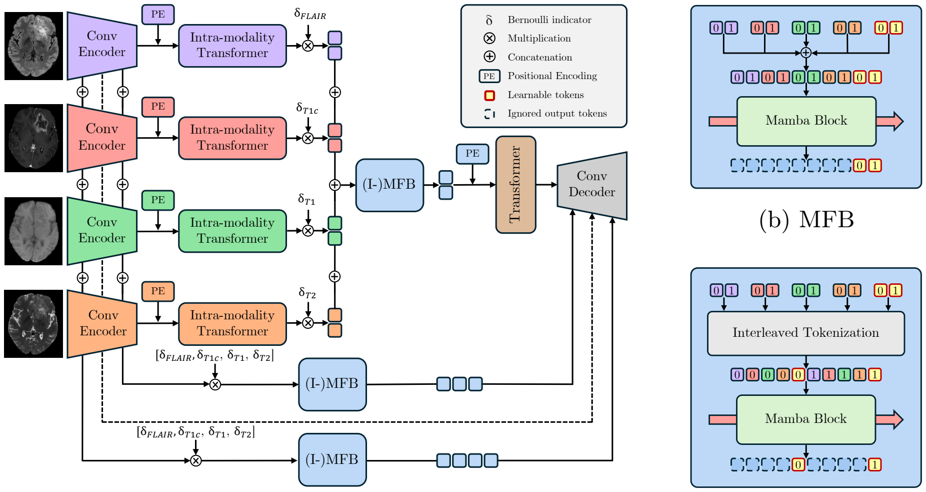
Contributions. We provide *i)* the very first extensive benchmarking of brain tumor segmentation algorithms under missing modality conditions on BraTS2023 dataset, *ii)* introduce a specifically designed Interleaved Mamba Fusion Block for effective multimodal fusion in the presence of missing image modalities, *iii)* and demonstrate its performance superiority while keeping the number of parameters and GFLOPs contained w.r.t. state-of-the-art models.

2 Method

2.1 Preliminaries

State-Space Model (SSM) is a mathematical framework used to represent dynamic systems wherein the input is mapped to an output with the same dimensionality through an N-dimensional latent state. The Mamba architecture builds on structured SSMs to manage long sequences effectively, imposing a structured constraint on its state transition matrix following the HiPPO theory [11] to boost memory retention and using a selection mechanism to focus on the most relevant information. This enhancement, combined with an efficient hardware-aware parallel algorithm, makes Mamba well-suited for effective and computationally efficient long-sequence modeling, with subquadratic complexity, by selectively propagating or discarding information along the sequence in an input-dependent manner. Moreover, given that flattened 3D volumes result in extremely long sequences, Mamba presents a promising solution for modeling long-range interactions, similar to transformer architectures, while avoiding the quadratic time complexity associated with self-attention mechanisms. Its

[4] BraTS2024 does not include a dedicated track for adult glioma segmentation, and BraTS2025 had not been released at the time of writing this paper.



(a) IM-Fuse

(b) MFB

(c) I-MFB

Fig. 1: (a) is an overview of our framework IM-Fuse (Incomplete Modality Fusion), (b) represents our Mamba Fusion Block (MFB) where learnable tokens are concatenated, and (c) depicts its interleaved version (Interleaved-MFB or I-MFB) where modality tokens and learnable parameters are alternately arranged.

efficient design weaves together linear projections and convolutions, making it well-suited for tasks involving complex sequential data. For more details, readers are encouraged to consult the original Mamba publication [12].

2.2 IM-Fuse - Incomplete Modality Fuse

Our proposal leverages hybrid modality-specific encoders to extract representations from each modality, Mamba to integrate multimodal features, a multimodal Transformer to capture long-range dependencies, and a convolutional decoder for reconstruction (Fig. 1a). The encoder-decoder structure follows 3D U-Net [8].

Given a multimodal input $X = \{\mathrm{X}_{\mathrm{FLAIR}}, \mathrm{X}_{\mathrm{T1c}}, \mathrm{X}_{\mathrm{T1}}, \mathrm{X}_{\mathrm{T2}}\}$, we define each modality-specific image as $\mathrm{X}_{\mathrm{m}} \in \mathbb{R}^{H \times W \times D}$, where $m \in \{\mathrm{FLAIR, T1, T1c, T2}\}$ represents the set of imaging modalities for each sample and $H \times W \times D$ denotes the spatial dimensions of the 3D medical volume.

Hybrid Modality-specific Encoder. Each modality m is first processed by a corresponding modality-specific convolutional encoder to extract high-level feature maps $\mathrm{F}_m^{\mathrm{local}} \in \mathbb{R}^{C \times \frac{H}{2^{(l-1)}} \times \frac{W}{2^{(l-1)}} \times \frac{D}{2^{(l-1)}}}$, where C represents the dimension of the feature channel and l denotes the number of stages in the convolutional encoder. The extracted feature maps are subsequently flattened into a 1D sequence, projected into a token space via a linear transformation, and a learnable positional embedding P_m is integrated to the projected features:

$$\mathrm{F}_m^{\mathrm{token}} = \mathrm{F}_m^{\mathrm{local}} \mathrm{W}_m + \mathrm{P}_m. \tag{1}$$

The resulting representation, $\mathrm{F}_m^{\text{token}}$, is fed to an intra-modal Transformer that models long-range dependencies within each modality, producing a context-aware global feature representation, $\mathrm{F}_m^{\text{global}} \in \mathbb{R}^{C\times \frac{H\times W\times D}{2^{3(l-1)}}}$, defined as follows:

$$\mathrm{F}_m^{\text{global}} = \mathrm{FFN}_m(\mathrm{LN}(z)) + \mathrm{z}, \qquad \mathrm{z} = \mathrm{MSA}_m(\mathrm{LN}(\mathrm{F}_m^{\text{token}})) + \mathrm{F}_m^{\text{token}}, \tag{2}$$

where the Multi-Head Self-Attention (MSA) enables the model to capture relationships between tokens through parallel attention heads, and the Feed-Forward Network (FFN) is a two-layer perceptron with a GELU activation.

Mamba as a Fusion Mechanism. Following [38], we introduce a Bernoulli indicator $\delta_m \in \{0, 1\}$ to simulate missing modalities during training. It is set to one if the modality m is present, zero otherwise. The corresponding embedding of each modality is simply multiplied by such a Bernoulli indicator $\delta_m F_m^{\text{global}}$. As a result, when a modality is missing, its corresponding feature vector is replaced by a zero vector and the multimodal representation will contain sparse information. To ensure consistent feature propagation, the Bernoulli indicator δ_m is applied also to the modality-specific features obtained from unimodal encoders at the i-th stage $\delta_m \mathrm{F}_m^{\text{local}_i}$, where $i = \{0, \dots, l-1\}$ and $\mathrm{F}_m^{\text{local}_i} \in \mathbb{R}^{C\times \frac{H}{2^{(i-1)}}\times \frac{W}{2^{(i-1)}}\times \frac{D}{2^{(i-1)}}}$, later integrated in the convolution decoder as skip connections.

Consequently, the sparse tensors propagate through both the skip connections and the bottleneck of the architecture, thereby negatively impacting performance. To address this issue, we leverage the long-sequence modeling capabilities and selection mechanism of Mamba to effectively integrate multimodal information while robustly handling missing data in both the skip connections and the bottleneck. Specifically, the long-sequence modeling capability facilitates the handling of both intra- and inter-modality interactions, whereas the selection mechanism allows the model to effectively disregard absent modalities. Thus, in this work, we introduce the Mamba Fusion Block (MFB), which accepts as input the tensors corresponding to the tokenized embeddings of the image modalities, each of dimensionality $\mathbb{R}^{P^3\times C}$, and produces an output $F_i^{\text{fused}} \in \mathbb{R}^{P^3\times C}$ that represents the fused representation. To achieve this, our MFB concatenates the modalities on a token-wise basis and appends a set of learnable tokens $K \in \mathbb{R}^{P^3\times C}$. Subsequently, a Mamba block processes the sequence from the first token of the modalities to the final learnable token, and only the output corresponding to the learnable tokens is propagated to subsequent layers, thereby obtaining the fused representation (Fig. 1b).

However, this approach may result in performance degradation if the combined number of modality tokens and learnable parameters is in the order of millions, which can occur in the case of large skip connections. In order to address this issue, we propose an interleaved concatenation strategy that gives rise to the Interleaved Mamba Fusion Block (I-MFB), wherein the modality tokens and learnable parameters are arranged alternately (Fig. 1c). Such an approach ensures that whenever Mamba generates a prediction for a learnable parameter token, the last four elements in its receptive field correspond to the modality tokens associated with the same supervoxel.

Multimodal Transformer. As done with the modality-specific feature maps, the fused representation obtained at the bottleneck by means of our Mamba Fusion Block, $\mathrm{F}^{\mathrm{fused}_l}$, is summed with a learnable positional embedding P_b and fed to a Transformer module, which, again, is intended to model long-range dependencies. The output of this module, $\mathrm{F}^{\mathrm{global}}$, is reshaped into feature maps and projected back into the convolutional space.

Convolutional Decoder. The convolutional decoder, designed symmetrically to the convolutional encoder, processes the fused representation, $\mathrm{F}^{\mathrm{global}}$, and the fused skip connections, $\mathrm{F}^{\mathrm{fused}_i}$, to progressively refine spatial resolution. By reconstructing the final segmentation mask from the high-level latent representation, it preserves low-level spatial details, enhancing segmentation performance.

Loss Function. Following [38], to ensure that each convolutional encoder produces detailed high-level feature representations, we employ a shared-weight decoder that independently performs tumor segmentation on the output of each modality-specific encoder, without relying on information from other modalities. The shared-weight decoder maintains the same architecture as the convolutional decoder. In addition, we interpolate the feature maps at each stage of the convolutional decoder to generate more accurate segmentation masks. We define the overall loss function as follows:

$$\mathcal{L}_{\mathrm{total}} = \sum_{i \in M} \mathcal{L}_i^{\mathrm{encoder}} + \sum_{i=1}^{l-1} \mathcal{L}_i^{\mathrm{decoder}} + \mathcal{L}^{\mathrm{output}}, \tag{3}$$

where $\mathcal{L}$ is jointly the Dice loss and weighted cross-entropy loss to handle the unbalanced object sizes in multi-class segmentation.

3 Experiments

Implementation Details. Our encoder architecture is based on a 3D U-Net and comprises five stages, each consisting of three sequential blocks. Each block includes a group normalization layer, a ReLU activation function, and a convolutional layer with a kernel size of 3. The first convolutional layer within each block has a stride of 2 to downsample the feature maps and doubles the number of feature channels. The first stage differs slightly, incorporating two blocks preceded by an initial convolutional layer that sets the number of feature channels to 8. Our method was implemented using Torch 2.5, and all models were trained on NVIDIA L40S GPUs with 48GB of memory each. In our network architecture, an (I-)MFB was integrated within the bottleneck and employed to fuse each of the five skip connections. The input dimensions for each image modality were set to 128 × 128 × 128 voxels, and the batch size was fixed at 2. The preprocessing and data augmentation pipeline was identical to that utilized by mmFormer [38]. The RAdam optimizer was employed, and a learning rate scheduler that progressively multiplies the learning rate by $(1 - \mathrm{epoch}/\mathrm{max\_epoch})^{0.9}$ during training, starting with an initial learning rate of 2×10^{-4}. We train our model for 1,000

Table 1: DSC% (↑) comparison across different missing modalities on BraTS2023 [2]. Present and missing modalities are denoted by ● and ○, respectively. Our proposal is identified with †. Best and second best results in **bold** and underline, respectively.

	Model	●F ○T1 ○T1c ○T2	○F ●T1 ○T1c ○T2	○F ○T1 ●T1c ○T2	○F ○T1 ○T1c ●T2	●F ●T1 ○T1c ○T2	●F ○T1 ●T1c ○T2	●F ○T1 ○T1c ●T2	○F ●T1 ●T1c ○T2	○F ●T1 ○T1c ●T2	○F ○T1 ●T1c ●T2	●F ●T1 ●T1c ○T2	●F ●T1 ○T1c ●T2	●F ○T1 ●T1c ●T2	○F ●T1 ●T1c ●T2	●F ●T1 ●T1c ●T2	**Avg.**
Enhancing Tumor	U-HVED	39.9	21.5	76.6	40.1	43.9	76.5	47.5	73.3	36.1	78.2	78.5	47.6	78.2	79.2	79.3	59.8
	Rob.Seg	50.3	44.8	80.2	50.5	55.4	82.7	57.4	81.7	53.9	82.4	83.7	59.0	82.8	82.9	83.6	68.8
	mmForm.	58.8	54.7	**84.1**	58.3	62.5	84.9	63.9	84.7	62.5	84.7	84.8	66.0	84.2	**85.9**	84.7	73.6
	SFusion	52.2	48.9	82.2	54.3	57.4	83.9	59.0	83.6	57.1	83.8	83.8	60.1	83.9	84.0	84.0	70.6
	ShaSpec	53.3	49.1	80.5	52.4	57.4	81.9	58.0	81.7	56.1	81.9	82.4	59.6	82.1	82.4	82.4	69.2
	M^3AE	56.7	56.0	82.5	58.8	60.7	**85.6**	61.2	**84.9**	60.6	**85.8**	**85.9**	62.3	**85.7**	85.1	85.6	73.2
	M^3FeCon	53.2	53.8	82.8	56.6	58.0	83.5	60.4	84.3	61.2	83.9	84.0	62.4	84.0	84.4	84.2	71.8
	IM-Fuse †	**59.5**	**56.3**	83.5	**59.6**	**63.9**	85.0	**64.3**	84.5	**63.6**	84.9	85.1	**67.0**	85.2	85.6	**85.8**	**74.3**
Tumor Core	U-HVED	58.9	47.1	81.9	61.7	68.0	81.9	66.5	84.9	60.2	84.8	84.4	68.9	84.4	85.3	86.0	73.7
	Rob.Seg	70.2	66.9	85.9	69.8	76.3	89.6	76.0	87.7	73.9	88.4	90.3	78.1	90.0	89.6	90.5	81.5
	mmForm.	78.3	73.6	89.2	74.5	80.6	90.7	**80.0**	90.1	77.5	90.6	90.9	80.8	91.0	90.8	91.0	84.7
	SFusion	74.0	70.4	86.7	74.3	77.4	88.6	77.8	88.4	76.5	89.4	89.1	78.5	89.3	89.5	89.5	82.6
	ShaSpec	74.3	71.5	87.8	72.6	78.1	89.7	77.3	89.3	76.2	89.6	90.3	79.1	90.4	90.0	90.7	82.8
	M^3AE	76.8	**75.9**	89.9	**77.9**	79.9	90.9	79.2	90.5	78.9	90.8	91.2	79.9	**91.5**	91.0	**91.5**	85.1
	M^3FeCon	72.3	74.1	90.1	75.9	77.1	90.9	79.1	90.6	**79.5**	90.7	91.2	80.9	91.2	91.4	91.1	84.4
	IM-Fuse †	**78.8**	75.4	**90.5**	76.5	**80.9**	**91.4**	79.9	**91.2**	79.1	**91.2**	**91.6**	**81.4**	91.3	**91.5**	**91.5**	**85.5**
Whole Tumor	U-HVED	82.8	71.2	63.5	80.1	86.8	86.8	87.9	79.0	83.9	85.1	88.4	89.3	90.0	86.8	90.8	83.5
	Rob.Seg	88.7	74.7	76.1	85.3	90.6	91.2	90.8	79.7	87.3	89.3	91.6	91.4	92.0	88.7	92.2	87.3
	mmForm.	91.4	82.8	**83.7**	88.5	92.2	92.7	91.3	**85.5**	89.8	90.1	**92.8**	92.5	93.0	**90.5**	93.0	90.0
	SFusion	89.1	78.5	77.6	87.0	90.8	91.2	91.3	81.3	88.2	88.7	91.7	91.6	92.1	88.8	92.2	88.0
	ShaSpec	91.0	79.9	79.5	86.9	91.9	92.3	92.2	82.7	88.3	88.7	92.6	92.5	92.9	89.2	93.0	88.8
	M^3AE	91.5	81.7	82.5	88.5	91.9	92.5	92.2	83.6	89.2	89.7	92.6	92.1	93.0	90.0	92.9	89.6
	M^3FeCon	87.7	81.2	81.2	88.5	89.4	90.0	92.1	83.9	89.8	89.5	90.5	92.7	92.6	90.1	93.1	88.8
	IM-Fuse †	**91.8**	**83.0**	**83.7**	**88.7**	**92.4**	**92.8**	**92.6**	**85.5**	**90.1**	**90.2**	**92.8**	**93.0**	**93.1**	**90.5**	**93.3**	**90.2**

epochs. We split the dataset[5] into 70% for training, 10% for validation, and 20% for testing, and the model selected for evaluation on the test set was the one that achieved the highest metric on the validation set.

Comparison with the State-of-the-Art. We retrained and compared the most prominent methods for brain tumor segmentation under missing modalities conditions, including U-HVED, RobustSeg, mmFormer, SFusion, ShaSpec, M^3AE, and M^3FeCon, alongside our proposed IM-Fuse, using the BraTS2023 dataset. For each model, we employed the same preprocessing, augmentation, optimizer, scheduler, and hyperparameters as described in their respective original papers, with the exception of the number of iterations, which were scaled to ensure an equivalent number of epochs as in the original studies due to the increased number of training samples. Additionally, for each model, we evaluated on the test set the version that achieved the highest metric on the validation set, thereby minimizing the risk of selecting an overfitted model. Results are presented in Tab. 1 for different tumor types evaluated across all 15 possible combinations of modality availability. Results indicate that our proposed

[5] Data splits are available at `https://github.com/AImageLab-zip/IM-Fuse`.

Table 2: Ablative study on the IM-Fuse fusion block. The first column represents the adopted fusion block, while the others report the DSC% metric across all the missing modalities scenarios on the BraTS2023 classes, ET, TC, and WT. ♣ denotes that the fusion block is applied to the bottleneck and skip connections simply concatenating the different modalities.

Fusion	ET	TC	WT	Avg.
MFB	53.42	72.61	76.10	67.38
I-MFB	**74.30**	**85.53**	**90.23**	**83.35**
MFB ♣	73.14	84.11	88.20	81.82
I-MFB ♣	73.26	84.27	89.09	82.20

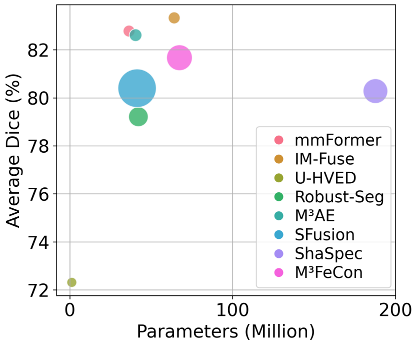


Fig. 2: Deployment model size and DSCs across all the missing modalities and tumor classes on BraTS2023. Larger circles → higher GFLOPS.

method, IM-Fuse, surpasses state-of-the-art architectures while maintaining an average computational complexity and parameter count contained, as illustrated in Fig. 2. In particular, our model excels at whole tumor segmentation, demonstrating superior performance in delineating the entire tumor region. All models achieved an average improvement of 8 Dice points by training on BraTS2023 compared to the performance obtained with BraTS2018 [28,31,38]. Furthermore, the relative performance among the models under examination changed as a result of the increased sample size in BraTS2023. Notably, mmFormer—employing a transformer architecture—now surpasses its convolutional-based competitors on BraTS2023. Indeed, transformers are known for requiring more data than convolution-based architectures and can benefit from the larger number of data samples available in the newer dataset version.

Ablation Study and Visualization. To demonstrate the effectiveness of the placement and design of I-MFB, we compared four different configurations on BraTS2023: I-MFB against MFB, each applied only in the bottleneck or in both bottleneck and skip connections. The experiments presented in Tab. 2 show that the proposed interleaved tokenization improves performance, and incorporating the I-MFB in the skip connections further enhances overall segmentation performance. MFB in the skip connections leads to performance degradation, but it is expected as the number of tokens in the skip connection is in the order of millions of tokens. Finally, visualization results are reported in Fig. 3 for different missing modality scenarios. According to quantitative results of Tab. 1, when T1c is missing, the enhancing tumor class (violet) is poorly identified.

4 Conclusion

In this study, we conduct a comprehensive comparison of the most prominent models in the field of brain tumor segmentation under missing modality conditions using the BraTS2023 dataset, which comprises 1,251 volumes—signifi-

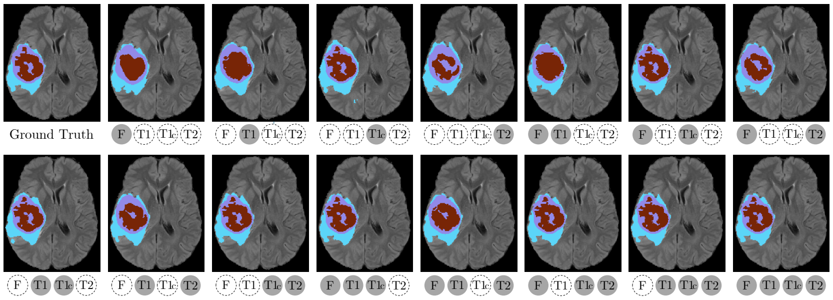


Fig. 3: Visualization of IM-Fuse segmentation results under missing modality scenarios. Enhancing tumor (violet), necrotic tissue (red), and edema (light blue).

cantly increasing the 285 volumes of its 2018 counterpart, where all the models were initially developed. Our analysis demonstrates that transformer-based architectures, which previously struggled with the smaller BraTS2018 dataset, now outperform purely convolutional models on BraTS2023 due to the increased data availability. Furthermore, we propose a Mamba-based architecture that achieves competitive performance compared to the state-of-the-art on BraTS2023.

Acknowledgments. This project has received funding from Fondazione di Modena, through the FAR 2024 (E93C24002080007), and by MUR, under the NRRP "Fit4MedRob-Fit for Medical Robotics" (PNC0000007) and the PRIN 2022 "AIDA: explAinable multImodal Deep learning for personAlized oncology" (20228MZFAA).

Disclosure of Interests. The authors have no conflicts of interest to declare.

References

1. Azad, R., Khosravi, N., Merhof, D.: SMU-Net: Style matching U-Net for brain tumor segmentation with missing modalities. In: MIDL (2022)
2. Baid, U., et al. (eds.): Brain Tumor Segmentation, and Cross-Modality Domain Adaptation for Medical Image Segmentation, Lecture Notes in Computer Science, vol. 14669 (2023)
3. Bakas, S., 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)
4. Bolelli, F., et al.: Segmenting the Inferior Alveolar Canal in CBCTs Volumes: the ToothFairy Challenge. IEEE TMI **44**(4) (2024)
5. Bolelli, F., et al.: Segmenting Maxillofacial Structures in CBCT Volumes. In: CVPR (2025)
6. Chen, C., et al.: Robust Multimodal Brain Tumor Segmentation via Feature Disentanglement and Gated Fusion. In: MICCAI (2019)
7. Chen, C., et al.: Learning With Privileged Multimodal Knowledge for Unimodal Segmentation. IEEE TMI **41**(3) (2021)
8. Çiçek, Ö., et al.: 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In: MICCAI (2016)

9. Ding, Y., Yu, X., Yang, Y.: RFNet: Region-Aware Fusion Network for Incomplete Multi-Modal Brain Tumor Segmentation. In: CVPR (2021)
10. Dorent, R., et al.: Hetero-Modal Variational Encoder-Decoder for Joint Modality Completion and Segmentation. In: MICCAI (2019)
11. Gu, A., et al.: HiPPO: Recurrent Memory with Optimal Polynomial Projections. In: NIPS. vol. 33 (2020)
12. Gu, A., et al.: Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv preprint arXiv:2312.00752 (2023)
13. Havaei, M., Guizard, N., Chapados, N., Bengio, Y.: HeMIS: Hetero-Modal Image Segmentation. In: MICCAI (2016)
14. Iv, M., et al.: Current Clinical State of Advanced Magnetic Resonance Imaging for Brain Tumor Diagnosis and Follow Up. In: Seminars in Roentgenology. vol. 53 (2018)
15. Jiang, Y., Shen, Y.: M4oE: A Foundation Model for Medical Multimodal Image Segmentation with Mixture of Experts. In: MICCAI (2024)
16. Karimijafarbigloo, S., et al.: MMCFormer: Missing Modality Compensation Transformer for Brain Tumor Segmentation. In: MIDL (2024)
17. Konwer, A., et al.: Enhancing Modality-Agnostic Representations via Meta-Learning for Brain Tumor Segmentation. In: CVPR (2023)
18. Lee, Y.L., et al.: Multimodal Prompting with Missing Modalities for Visual Recognition. In: CVPR (2023)
19. Liu, H., et al.: M3AE: Multimodal Representation Learning for Brain Tumor Segmentation with Missing Modalities. In: AAAI. vol. 37 (2023)
20. Liu, Z., Wei, J., Li, R., Zhou, J.: SFusion: Self-attention Based N-to-One Multimodal Fusion Block. In: MICCAI (2023)
21. Lumetti, L., et al.: Enhancing Patch-Based Learning for the Segmentation of the Mandibular Canal. IEEE Access (2024)
22. Lumetti, L., et al.: Taming Mambas for 3D Medical Image Segmentation. IEEE Access (2025)
23. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1) (2024)
24. Pipoli, V., et al.: Semantically Conditioned Prompts for Visual Recognition under Missing Modality Scenarios. In: WACV (2025)
25. Qiu, Y., Zhao, Z., Yao, H., Chen, D., Wang, Z.: Modal-aware Visual Prompting for Incomplete Multi-modal Brain Tumor Segmentation. In: MM (2023)
26. Qiu, Y., et al.: Scratch Each Other's Back: Incomplete Multi-Modal Brain Tumor Segmentation via Category Aware Group Self-Support Learning. In: CVPR (2023)
27. Shen, Y., Gao, M.: Brain Tumor Segmentation on MRI with Missing Modalities. In: IPMI (2019)
28. Shi, J., et al.: MFTrans: Modality-Masked Fusion Transformer for Incomplete Multi-Modality Brain Tumor Segmentation. IEEE J-BHI **28**(1) (2023)
29. Shi, J., et al.: PASSION: Towards Effective Incomplete Multi-Modal Medical Image Segmentation with Imbalanced Missing Rates. In: MM (2024)
30. Vadacchino, S., et al.: HAD-Net: A Hierarchical Adversarial Knowledge Distillation Network for Improved Enhanced Tumour Segmentation Without Post-Contrast Images. In: MIDL (2021)
31. Wang, H., et al.: Multi-Modal Learning With Missing Modality via Shared-Specific Feature Modelling. In: CVPR (2023)
32. Wang, S., et al.: Prototype Knowledge Distillation for Medical Segmentation with Missing Modality. In: ICASSP (2023)

33. Wang, X., Li, Z., Huang, Y., Jiao, Y.: Multimodal medical image segmentation using multi-scale context-aware network. Neurocomputing **486** (2022)
34. Wang, Y., et al.: ACN: Adversarial Co-training Network for Brain Tumor Segmentation with Missing Modalities. In: MICCAI (2021)
35. Yang, Q., et al.: D 2-Net: Dual Disentanglement Network for Brain Tumor Segmentation With Missing Modalities. IEEE TMI **41**(10) (2022)
36. Zeng, Z., et al.: Missing as Masking: Arbitrary Cross-Modal Feature Reconstruction for Incomplete Multimodal Brain Tumor Segmentation. In: MICCAI (2024)
37. Zhang, Y., et al.: Modality-Aware Mutual Learning for Multi-modal Medical Image Segmentation. In: MICCAI (2021)
38. Zhang, Y., et al.: mmFormer: Multimodal Medical Transformer for Incomplete Multimodal Learning of Brain Tumor Segmentation. In: MICCAI (2022)
39. Zhou, T., Canu, S., Vera, P., Ruan, S.: Latent Correlation Representation Learning for Brain Tumor Segmentation With Missing MRI Modalities. IEEE TMI **30** (2021)
40. Zhou, T., Ruan, S., Hu, H.: A literature survey of MR-based brain tumor segmentation with missing modalities. CMIG **104** (2023)
41. Zhu, Y., et al.: DRM-VAE: A Dual Residual Multi Variational Auto-Encoder for Brain Tumor Segmentation with Missing Modalities. In: ICECE (2021)