Incomplete Multi-modal Brain Tumor Segmentation via Learnable Sorting State Space Model

Supplementary Material

7. Appendix

In the appendix, we present a comprehensive analysis covering the following aspects: limitations of our approach, extended comparison on BraTS2020, model complexity analysis, comparison with CNN-based and Transformer-based architectures, the necessity of SSMs-based design, an ablation study on permutation matrix generation for long sequences, an analysis of Mamba's computation graph, and comparison on BraTS2018.

7.1. Limitation

While LS3M demonstrates promising performance, it encounters limitations:

1. **Permutation Matrix Assumptions:** The predicted permutation matrix P inherently assumes a strict structure where each row and column contains exactly one '1', with all other entries being '0s'. Ideal conditions would require the temperature parameter τ to approach zero and the iteration times n to be infinitely large to achieve a perfect permutation matrix. However, practical constraints lead us to set $\tau = 1$ and $n = 10$, resulting in a matrix P that only approximates this ideal, with minimal but non-zero values where zeros are expected.
2. **Dataset Limitations:** Our evaluations of LS3M are conducted using BraTS2018 and BraTS2020, with the former being a subset of the latter. Ideally, additional datasets unaffiliated with the BraTS series would provide a better validation of LS3M's capabilities. Unfortunately, suitable alternatives are scarce. Several factors contribute to the scarcity of suitable alternatives:
   - **Specific Task Requirements:** Our task demands multiple imaging modalities and a single segmentation mask per patient, with modality image dimensions being integer multiples of $16 \times 16 \times 16$ to ensure compatibility with multi-scale processing. These stringent requirements significantly restrict the availability of compatible datasets.
   - **Issues with Existing Datasets:**
     - **ISLES (Ischemic Stroke Lesion Segmentation [22]):** This dataset includes multi-modal images such as FLAIR, DWI, and ADC. However, the FLAIR and DWI modalities are not well-registered, and the volume size is too small for our task.
     - **CHAOS (Combined (CT-MR) Healthy Abdominal Organ Segmentation [29]) and AMOS (Abdominal Multi-Organ Segmentation [26]):** These datasets provide CT and MRI modalities but only one modality per patient, failing to meet the multi-modal requirement.
     - **MSD (Medical Segmentation Decathlon [2]):** While this dataset includes CT and MRI modalities across different organs, most of these are single-modality datasets per patient, except for the brain dataset, which is derived from BraTS.

To ensure a fair comparison, we align with prior works such as RFNet [9], mmFormer [62], and M3AE [36], which are among the most prominent works in this domain. These methods have consistently used the BraTS2018 and BraTS2020 datasets as benchmarks, which are considered standard for multi-modal brain tumor segmentation with missing modalities. Subsequent research has largely followed this precedent, further validating the rationality of using these datasets.

While we acknowledge the importance of diverse datasets for broader validation, our current evaluation leverages the most suitable and widely accepted datasets in this field. We aim to explore additional datasets that align with our specific task requirements in future work.

7.2. Extended Comparison on BraTS2020

In addition to the main experiments presented in Tab. 1, , which compared nine models across 15 different missing modality scenarios using the Dice coefficient (DSC), we provide an extended evaluation on BraTS2020 in this appendix. Here, we summarize key performance metrics DSC, 95% Hausdorff distance (HD95), sensitivity (Sens(%)), and specificity (Spec(%)) focusing on the average results across the 15 missing modality scenarios.

Moreover, we report the standard error (SE) and the 95% confidence interval bounds (LB and HB) for the results, computed using bootstrapping [42] with 10,000 resampling iterations. This provides insights into the robustness and reliability of the proposed model's performance, particularly for clinically relevant scenarios.

Table 4 compares our proposed LS3M model against other state-of-the-art methods across the Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET) regions. Unlike the original table, which focused solely on DSC for all 15 missing modality scenarios, this table presents a more comprehensive set of metrics, offering a deeper understanding of each model's segmentation quality. Our LS3M consistently outperforms the baselines across all metrics, especially in DSC and HD95, indicating superior segmentation accuracy and boundary precision, thus

demonstrating its effectiveness in clinical scenarios with missing modalities.

7.3. Comparison of Model Complexity

To assess the model complexity of our LS3M, we compare the number of parameters and GFLOPS with other existing models, as shown in Table 5. Compared to UNet-MFI [66], LS3M achieves 6% higher performance while requiring only 35% of the GFLOPS, demonstrating the efficiency of our global modeling designs, including the combination of SortP, the S3M block, and the Global Input Strategy (GIS).

Notably, mmFormer [62] employs Transformer blocks to model long-range dependencies, but only at a very low resolution, specifically $4 \times 4 \times 4$. While this reduces computational burden, it limits the ability to build detailed correlations, resulting in average DSC scores for WT, TC, and ET of around 82.92%, 74.90%, and 58.02%, respectively. In contrast, LS3M is capable of building long-range dependencies at higher resolutions ($20 \times 20 \times 20$ and $10 \times 10 \times 10$), leading to superior performance across all metrics (88.22%, 79.76%, and 63.60% for WT, TC, and ET, respectively).

The comparison results showcase that LS3M not only provides state-of-the-art segmentation accuracy but also maintains an efficient balance between model complexity and performance, highlighting its potential for practical applications.

7.4. Comparison with CNN-based and Transformer-based Architectures

To evaluate the superiority of our SSM-based architecture, we first simplify LS3M by excluding the SortP and S3M components, resulting in a purely CNN-based architecture. Next, we add a Transformer block in place of the SortP and S3M blocks to create a Transformer-based version. Finally, we scale the CNN-based architecture to match the same level of GFLOPs as our model for a fair comparison. The results are presented in Tab. 6.

The GFLOPs row shows the computational complexity of each model. Although Transformer-based architectures can capture long-range dependencies, they incur 35% higher GFLOPS compared to our SSMs-based architecture. In terms of segmentation accuracy, our architecture outperforms the others across all tumor regions: Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET). Notably, the scaled CNN model, with GFLOPs similar to ours, still falls short of our model's performance, emphasizing the effectiveness of our design.

7.5. Necessity of SSMs-based Design

Accurate brain tumor segmentation in incomplete multi-modal scenarios is challenging due to the lack of sufficient information, making it difficult to capture tumor boundaries and spatial relationships effectively. Long-range modeling plays a critical role in bridging this gap by providing a more comprehensive representation that compensates for missing modalities and ensures robust segmentation. As illustrated in Sec. 7.3 and Sec. 7.4, while Transformer-based architectures can mitigate the challenges of long-range modeling, they exhibit significant limitations. These include high computational costs or the necessity to operate at very low resolutions, which limits their ability to capture fine-grained details effectively. For example, mmFormer [62] uses Transformer blocks only at low resolutions ($4 \times 4 \times 4$) to reduce computational load, but this compromises the ability to build detailed spatial correlations. In contrast, UNet-MFI [66] requires 499.52 GFLOPS. The high computational costs arise from the large sequence lengths inherent in 3D brain MRI data, resulting in quadratic increases in computation for high-resolution images.

Therefore, employing a more efficient approach to build long-range dependencies is critical. Our LS3M addresses these challenges by using SSMs-based designs, which efficiently model both local and long-range dependencies while maintaining lower computational overhead. This allows LS3M to operate at higher resolutions ($20 \times 20 \times 20$ and $10 \times 10 \times 10$) without incurring the prohibitive computational costs associated with Transformers, thus striking a balance between computational efficiency and segmentation accuracy.

7.6. Ablation Study on Permutation Matrix Generation for Long Sequences.

In our LS3M framework, we propose an efficient alternative to handle longer sequences for features at higher resolutions by using bilinear interpolation to scale the permutation matrix instead of employing SortNet. We conducted ablation experiments to compare our approach with three other methods. Specifically, we used SortNet as the first comparison model. Additionally, we used the permutation matrix from the lower resolution and scaled it using three different up-sampling techniques: Transposed Convolution (TransConv), Bilinear Interpolation with Corner Alignment (BI-CA), and Bilinear Interpolation without corner alignment (BI), which is our proposed method.

Table 7 presents the results of these ablation experiments. The comparison highlights that our method (Bi-AF) achieves the best overall performance for both WT and TC, with only a slight performance drop for ET compared to SortNet. Notably, our approach maintains a competitive balance between segmentation accuracy and computational simplicity, demonstrating its effectiveness in generating permutation matrices for higher-resolution features.

	Model	DSC↑				HD95↓				Sens↑				Spec↑			
		AVG	SE	LB	HB	AVG	SE	LB	HB	AVG	SE	LB	HB	AVG	SE	LB	HB
WT	RFNet	86.76	0.30	86.16	87.31	7.91	**0.33**	7.29	8.57	88.38	0.35	87.69	89.06	99.65	7.56e-5	99.64	99.67
	mmFormer	82.92	0.45	82.10	83.83	11.13	0.43	10.30	11.97	79.80	0.53	78.72	80.84	**99.68**	**7.52e-5**	**99.67**	**99.70**
	GSS	87.00	0.26	86.48	87.52	7.58	0.43	6.77	8.45	89.71	**0.29**	89.12	90.26	99.60	9.63e-5	99.58	99.61
	Ours	**88.22**	**0.26**	**87.68**	**88.71**	**7.10**	0.38	**6.37**	**7.87**	**90.19**	0.30	**89.58**	**90.76**	99.65	9.69e-5	99.65	99.66
TC	RFNet	77.39	0.57	76.62	78.85	8.86	**0.38**	8.15	9.63	82.80	0.58	81.67	83.93	99.77	7.34e-5	99.76	99.78
	mmFormer	74.90	0.66	73.61	76.19	12.22	0.39	11.48	13.02	66.72	0.71	65.30	68.10	99.81	6.69e-5	99.80	99.82
	GSS	78.45	0.56	77.33	79.57	8.24	0.54	7.20	9.31	82.99	0.58	81.80	83.99	99.81	5.92e-5	99.80	99.82
	Ours	**79.76**	**0.55**	**78.63**	**80.76**	**7.88**	0.42	**7.07**	**8.69**	**85.12**	**0.55**	**84.06**	**86.19**	**99.83**	**5.81e-5**	**99.82**	**99.84**
ET	RFNet	60.93	0.85	59.28	62.58	7.52	0.39	6.78	8.30	74.45	**0.71**	73.06	75.84	99.73	9.24e-5	99.72	99.75
	mmFormer	58.02	0.83	56.41	59.65	9.00	0.42	8.21	9.84	55.68	0.92	53.91	57.51	**99.88**	**3.71e-5**	**99.88**	**99.89**
	GSS	61.15	0.69	59.79	62.47	7.21	0.47	6.31	8.16	74.98	**0.71**	73.53	76.37	99.81	6.02e-5	99.80	99.83
	Ours	**63.60**	**0.67**	**62.31**	**64.93**	**6.05**	**0.39**	**5.32**	**6.86**	**76.34**	0.73	**74.90**	**77.73**	99.82	6.02e-5	99.81	99.84

Table 4. Comparison of models across different metrics (DSC↑, HD95↓, Sens↑, Spec↑).

	HeMiS	U-HVED	RFNet	UNet-MFI	mmFormer	M3AE	Ours
Param(M)	0.57	1.25	8.98	34.12	36.56	40.42	33.38
GFLOPS	2.27	4.58	102.28	499.52	30.23	36.14	171.58
WT	69.88	62.76	86.76	82.21	82.92	86.25	**88.22**
TC	52.76	43.53	77.39	74.01	74.90	77.56	**79.76**
ET	42.69	30.64	60.93	57.52	58.02	61.30	**63.60**

Table 5. Comparison with SOTA methods on model complexity on BraTS2020.

	CNN-based	Transformer-based	Ours	CNN-GFLOPs
GFLOPs	115.30	236.39	171.58	176.41
WT	86.23	**88.39**	88.22	87.23
TC	77.06	79.47	**79.76**	78.75
ET	60.35	63.26	**63.60**	62.13

Table 6. Comparison of CNN-based, Transformer-based, and Our SSMs-based models.

	SortNet	TransConv	BI-CA	BI (Ours)
WT	88.15	87.91	88.12	**88.22**
TC	79.64	79.13	79.55	**79.76**
ET	**63.69**	62.53	62.62	63.60

Table 7. Effect of permutation matrix generation for long sequences.

7.7. Comparison for RGB-depth task on NYUv2.

To explore the potential generalizability of our LS3M to natural image segmentation tasks, we follow DMRNet's setting [55], and adapt RFNet, mmFormer, and our model to RGB-depth task on NYUv2 [48]. We unify modality encoders using pretrained ResNet-50, while retaining core components and switching from 3D to 2D processing. As shown in Tab. 8, results show our superiority on the RGB-depth segmentation task.

7.8. Clarity of Permutation Matrix Generation.

To enhance the understanding of our permutation process, we provide a visualization using a toy example of array permutation. As illustrated in Fig. 6, for an $(L \times C)$ token sequence, SortNet generates an $(L \times L)$ matrix. Then, Gumbel-Softmax makes rows approximate one-hot vectors, and the Sinkhorn operator makes the matrix doubly-stochastic (rows and columns sum to 1).

Modalities		mIOU ↑			
RGB	Depth	RFNet	mmFormer	DMRNet	Ours
•	○	42.89	43.22	44.10	**45.01**
○	•	40.76	41.12	41.88	**43.02**
•	•	48.13	48.45	49.27	**49.93**
Average		43.92	44.26	45.08	**45.98**

Table 8. Performance comparison on NYUv2.

7.9. Analysis of Mamba's Computational Graph

We analyze the computational graph of the Mamba block to provide rationality for the inclusion of a channel attention block in our S3M and the concatenation for multi-modal sequences along the channel dimension for multi-modal fusion. Given the similarities between our Mamba-like block and the original Mamba block, the characteristics and behaviors of the two are expected to be analogous.

As depicted in Figure 7, the original Mamba block processes channel interactions primarily through '*Conv_1d*' and '*In_proj*' operations. The capability of these operations to model channel-wise dependencies is inherently limited by the constraints of convolutional and linear layers. To enhance the modeling of these dependencies, we incorporate a channel attention block within our S3M framework.

After concatenating multi-modal sequences along the channel dimension, the combined data is compressed into X_{dbl}', which is further split into SSM parameters Δ, B, C. This process translates information from the channel dimension into the state dimension. The operation '*torch.einsum('bdl,bnl,bdl-¿bdln', delta, B, u)*' facilitates the interaction between the channel dimensions of the multi-modal sequences and the state dimensions of the SSM parameters, culminating in an effective multi-modal fusion.

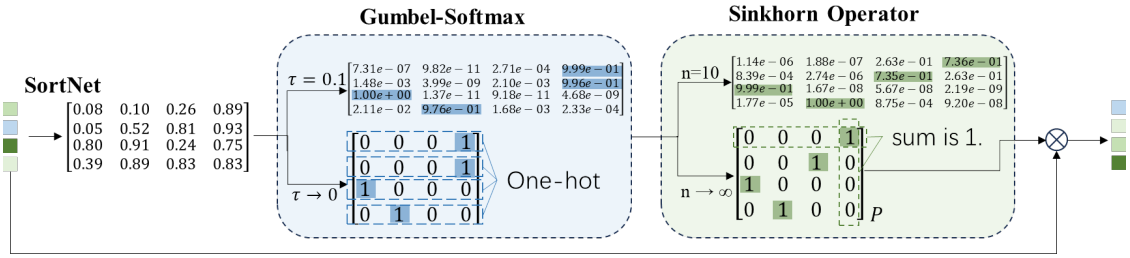

Figure 6. Permutation Matrix Generation.

7.10. Comparison on BraTS2018

We compare our method with five state-of-the-art methods on BraTS2018, employing a three-fold validation approach as detailed in Tab. 9. For each method, we provide the mean and standard error of the Dice score in varying scenarios with missing modalities. The results consistently indicate that our LS3M outperforms the compared state-of-the-art methods across all tumor categories, including whole tumor (WT), tumor core (TC), and enhancing tumor (ET).

The results in Tab. 9 highlight the superior performance of our method across various missing modality scenarios. Our approach achieves the highest Dice scores for all three tumor regions in most cases, reflecting its robustness in handling incomplete data. Additionally, the reduced standard errors in comparison to other methods indicate greater consistency in segmentation performance across folds. This advantage is attributed to our framework's designs to effectively leverage available modalities and model long-range dependencies, ensuring accurate and reliable segmentation results.

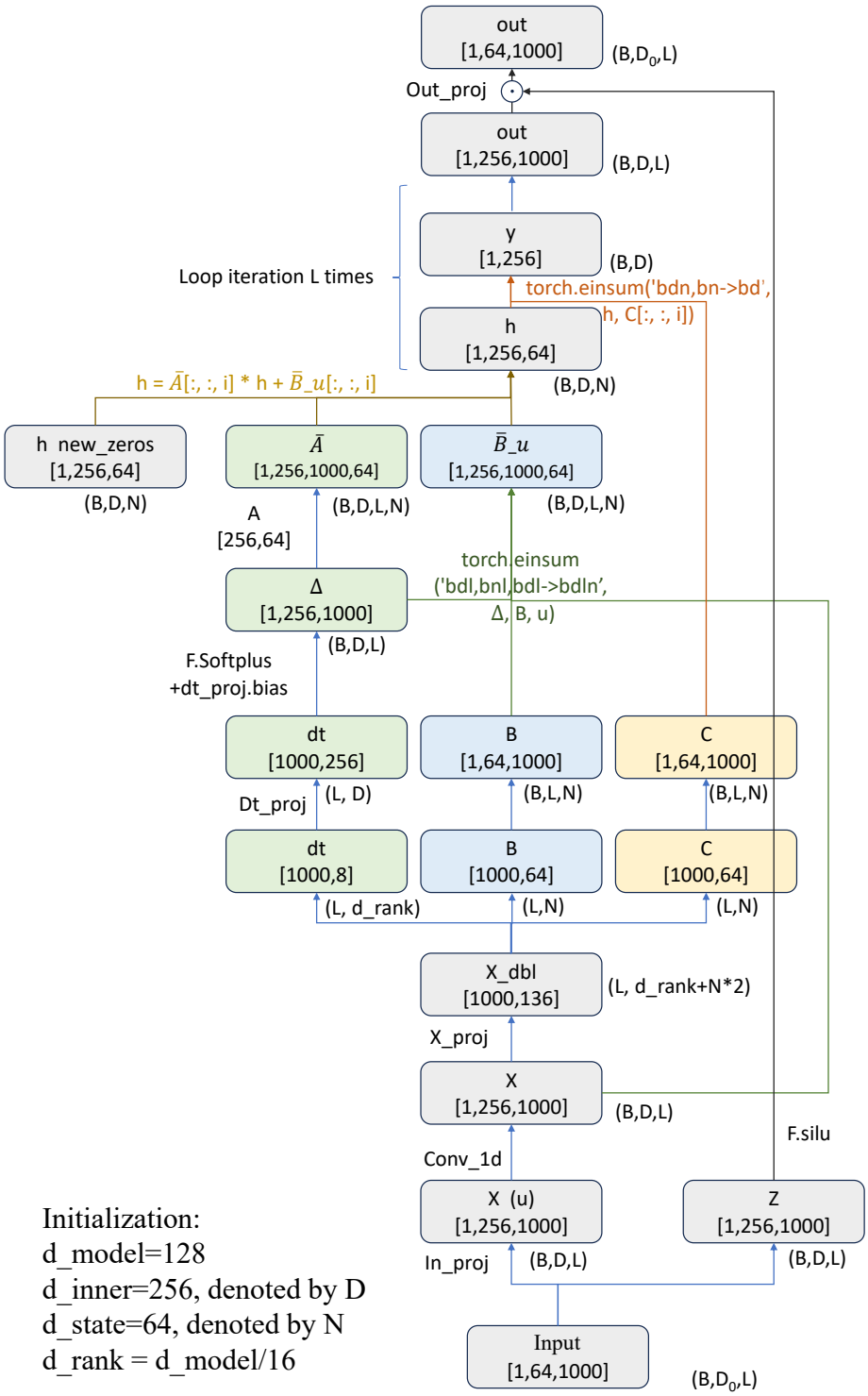

Figure 7. Mamba's Computational Graph.

M																	AVG
	FLAIR	•	○	○	○	•	•	•	○	○	○	•	•	•	○	•	
	T1ce	○	•	○	○	•	○	○	•	•	○	•	•	○	•	•	
	T1	○	○	•	○	○	•	○	•	○	•	•	○	•	•	•	
	T2	○	○	○	•	○	○	•	○	•	•	○	•	•	•	•	
WT	UNet-MFI	84.06	65.90	68.80	85.23	90.23	89.88	92.75	75.46	88.40	89.47	90.97	93.27	93.82	89.71	93.81	86.12
		±0.14	±1.00	±0.55	±0.14	±0.20	±0.37	±0.36	±0.29	±0.06	±0.28	±0.39	±0.16	±0.19	±0.33	±0.05	±0.02
	RFNet	83.01	68.94	71.78	84.27	89.25	89.47	90.50	75.76	87.30	87.99	90.54	91.57	92.00	88.57	92.18	85.54
		±2.10	±0.64	±0.72	±0.65	±0.95	±0.85	±1.03	±0.51	±0.70	±0.66	±0.62	±0.74	±0.64	±0.78	±0.57	±0.76
	mmFormer	81.10	64.44	66.87	82.77	87.22	87.99	89.32	72.33	85.48	86.26	88.28	89.96	90.44	86.61	90.52	83.24
		±0.47	±0.68	±1.15	±0.03	±0.49	±0.40	±0.51	±0.38	±0.16	±0.21	±0.38	±0.43	±0.33	±0.29	±0.39	±0.27
	M3AE	83.48	65.67	70.54	79.68	90.69	89.39	89.04	75.69	85.64	87.28	91.35	92.03	92.55	88.53	92.10	83.98
		±3.03	±1.52	±1.48	±2.96	±1.97	±1.53	±2.48	±0.96	±2.63	±1.87	±1.13	±1.95	±1.67	±1.42	±1.39	±1.70
	ShapSpec	82.76	69.23	71.09	85.50	90.43	89.09	92.01	75.74	88.72	89.06	91.37	93.03	93.27	89.71	93.39	86.36
		±2.34	±1.26	±0.43	±0.59	±1.13	±0.75	±0.99	±0.72	±0.63	±0.40	±0.76	±0.75	±0.55	±0.64	±0.60	±0.77
	Ours	85.12	73.72	74.81	84.83	93.17	89.83	90.03	80.84	88.27	91.60	91.24	93.85	94.41	92.48	95.89	88.01
		±0.42	±0.54	±0.39	±0.36	±0.39	±0.42	±0.44	±0.42	±0.40	±0.40	±0.41	±0.43	±0.44	±0.43	±0.44	±0.42
TC	UNet-MFI	56.77	53.93	80.70	63.35	69.62	84.89	69.86	85.11	68.82	87.24	88.15	73.36	89.13	89.27	90.17	76.69
		±0.72	±0.89	±0.30	±0.70	±0.03	±0.27	±0.26	±0.21	±0.95	±0.39	±0.09	±0.35	±0.23	±0.56	±0.02	±0.29
	RFNet	60.88	57.42	77.81	63.89	71.58	85.99	71.40	80.37	70.21	85.83	86.13	74.21	87.56	86.20	87.90	76.43
		±1.98	±0.85	±1.06	±1.91	±1.05	±0.99	±0.87	±0.88	±0.81	±0.77	±0.88	±0.45	±0.81	±0.88	±0.87	±0.69
	mmFormer	52.60	52.69	77.61	61.18	66.22	82.07	66.25	81.10	65.87	84.20	84.79	69.58	85.78	85.63	86.47	73.47
		±1.18	±0.34	±0.08	±1.48	±0.31	±0.17	±1.35	±0.36	±1.20	±0.65	±0.07	±0.60	±0.40	±0.47	±0.08	±0.35
	M3AE	61.45	56.93	77.97	61.85	73.08	86.50	71.45	81.48	69.44	87.53	88.66	75.35	90.47	88.45	90.91	77.37
		±1.36	±1.09	±2.21	±0.91	±1.20	±1.55	±1.58	±1.88	±0.96	±1.43	±1.29	±1.05	±1.25	±1.35	±1.23	±0.93
	ShapSpec	60.71	58.56	78.22	64.80	73.34	86.65	72.58	81.72	70.81	87.54	88.03	75.77	89.88	87.99	90.26	77.79
		±2.33	±0.64	±0.91	±1.96	±1.27	±0.92	±1.30	±0.80	±1.10	±0.72	±0.97	±0.71	±0.75	±0.80	±0.82	±0.66
	Ours	63.29	60.15	81.76	63.00	75.62	87.20	73.16	85.59	71.26	86.49	89.55	79.39	90.23	89.90	91.55	79.28
		±1.02	±1.15	±1.16	±1.09	±1.08	±1.09	±1.08	±1.10	±1.07	±1.07	±1.07	±1.08	±1.06	±1.06	±1.08	±1.08
ET	UNet-MFI	37.69	25.23	74.22	41.37	41.27	77.88	44.86	76.67	42.87	78.48	78.04	46.42	78.41	78.98	78.96	60.16
		±0.66	±2.76	±1.00	±0.94	±0.93	±0.69	±1.06	±0.71	±0.91	±0.22	±0.61	±0.87	±0.29	±0.17	±0.43	±0.27
	RFNet	37.21	31.97	66.43	40.72	42.30	71.54	44.12	68.96	43.37	72.19	72.33	45.54	73.24	72.53	73.49	57.06
		±0.83	±0.60	±1.33	±1.04	±0.95	±0.62	±0.23	±0.76	±0.46	±0.78	±0.81	±0.31	±0.84	±0.84	±0.83	±0.51
	mmFormer	32.22	24.93	75.80	41.27	39.69	79.10	44.31	78.20	43.04	79.51	80.08	45.69	79.38	80.24	80.06	60.23
		±1.02	±1.14	±0.45	±0.94	±0.71	±0.43	±1.17	±0.30	±1.14	±0.12	±0.19	±0.69	±0.22	±0.04	±0.29	±0.28
	M3AE	40.50	33.50	68.13	41.40	46.40	75.28	46.62	71.51	44.55	75.04	76.32	48.65	77.33	76.87	77.60	60.96
		±1.35	±2.11	±2.39	±0.78	±1.74	±1.78	±0.61	±2.05	±0.51	±1.86	±1.63	±0.69	±1.44	±1.62	±1.55	±1.33
	ShapSpec	39.73	35.26	69.94	44.40	45.83	76.05	47.81	73.16	46.00	76.78	77.10	49.45	78.18	77.24	78.35	61.09
		±0.25	±0.62	±1.29	±1.42	±0.25	±0.75	±0.88	±1.09	±1.03	±0.89	±0.90	±0.23	±0.85	±0.93	±0.85	±0.24
	Ours	41.54	36.42	68.81	45.56	47.94	76.59	52.56	75.75	48.59	78.92	81.86	53.96	78.45	79.81	80.96	63.31
		±0.83	±0.44	±0.62	±0.75	±0.80	±0.83	±0.82	±0.82	±0.88	±0.89	±0.91	±0.90	±0.90	±0.91	±0.81	±0.81

Table 9. Comparison with five SOTA methods, including RFNet, UNet-MFI, mmFormer, M3AE, and ShaSpec on BraTS2018. The performance on whole tumor, tumor core, and enhancing tumor segmentation are evaluated by the dice scores (reported as Mean ± Standard Error). Red and blue indicate the $1st$ and $2nd$ ranks, respectively. Additionally, we use • to denote available modalities and ○ to denote missing modalities.

References

[1] Ryan Prescott Adams and Richard S Zemel. Ranking via sinkhorn propagation. *arXiv preprint arXiv:1106.1925*, 2011. 3

[2] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022. 1

[3] Reza Azad, Nika Khosravi, and Dorit Merhof. Smu-net: Style matching u-net for brain tumor segmentation with missing modalities. In *International Conference on Medical Imaging with Deep Learning*, pages 48–62. PMLR, 2022. 1, 2

[4] Ali Behrouz and Farnoosh Hashemi. Graph mamba: Towards learning on graphs with state space models. *arXiv preprint arXiv:2402.08678*, 2024. 3

[5] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision–ECCV 2022 Workshops*, pages 205–218. Springer, 2023. 2

[6] Cheng Chen, Qi Dou, Yueming Jin, Hao Chen, Jing Qin, and Pheng-Ann Heng. Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, pages 447–456. Springer, 2019. 2

[7] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2

[8] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024. 1

[9] Yuhang Ding, Xin Yu, and Yi Yang. Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3975–3984, 2021. 1, 2, 6, 7

[10] Yuhang Ding, Liulei Li, Wenguan Wang, and Yi Yang. Clustering propagation for universal medical image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2024. 2

[11] Reuben Dorent, Samuel Joutard, Marc Modat, Sébastien Ourselin, and Tom Vercauteren. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*, pages 74–82. Springer, 2019. 7

[12] Zhengcong Fei, Mingyuan Fan, Changqian Yu, and Junshi Huang. Scalable diffusion models with state space backbone. *arXiv preprint arXiv:2402.05608*, 2024. 3

[13] Jiannan Ge, Lingxi Xie, Hongtao Xie, Pandeng Li, Xiaopeng Zhang, Yongdong Zhang, and Qi Tian. Alignzeg: Mitigating objective misalignment for zero-shot semantic segmentation. In *European Conference on Computer Vision*, pages 142–161. Springer, 2024. 2

[14] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 1, 3

[15] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2021. 2

[16] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. *arXiv preprint arXiv:2402.15648*, 2024. 3

[17] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event*, pages 272–284. Springer, 2022. 2

[18] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022. 2

[19] Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. Hemis: Hetero-modal image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, pages 469–477. Springer, 2016. 7

[20] Xuanhua He, Ke Cao, Keyu Yan, Rui Li, Chengjun Xie, Jie Zhang, and Man Zhou. Pan-mamba: Effective pan-sharpening with state space model. *arXiv preprint arXiv:2402.12192*, 2024. 3

[21] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

[22] Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific data*, 9(1): 762, 2022. 1

[23] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5

[24] Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. Localmamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338*, 2024. 3, 4

[25] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2016. 3

[26] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022. 1

[27] Zeyu Jiang, Changxing Ding, Minfeng Liu, and Dacheng Tao. Two-stage cascaded u-net: 1st place solution to brats

challenge 2019 segmentation task. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 231–241. Springer, 2020. 2

[28] Sanaz Karimijafarbigloo, Reza Azad, Amirhossein Kazerouni, Saeed Ebadollahi, and Dorit Merhof. Mmcformer: Missing modality compensation transformer for brain tumor segmentation. In *Medical Imaging with Deep Learning*, pages 1144–1162. PMLR, 2024. 2

[29] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021. 1

[30] Jonghun Kim and Hyunjin Park. Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7604–7613, 2024. 1, 2

[31] Dongwook Lee, Won-Jin Moon, and Jong Chul Ye. Assessing the importance of magnetic resonance contrasts using collaborative generative adversarial networks. *Nature Machine Intelligence*, 2(1):34–42, 2020. 1, 2

[32] Ho Hin Lee, Shunxing Bao, Yuankai Huo, and Bennett A Landman. 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. *arXiv preprint arXiv:2209.15076*, 2022. 2

[33] Jiangyun Li, Wenxuan Wang, Chen Chen, Tianxiang Zhang, Sen Zha, Hong Yu, and Jing Wang. Transbtsv2: Wider instead of deeper transformer for medical image segmentation. *arXiv preprint arXiv:2201.12785*, 2022. 2

[34] Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2024. 2

[35] Dingkang Liang, Xin Zhou, Xinyu Wang, Xingkui Zhu, Wei Xu, Zhikang Zou, Xiaoqing Ye, and Xiang Bai. Pointmamba: A simple state space model for point cloud analysis. *arXiv preprint arXiv:2402.10739*, 2024. 3

[36] Hong Liu, Dong Wei, Donghuan Lu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. M3ae: Multimodal representation learning for brain tumor segmentation with missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1657–1665, 2023. 1, 2, 6, 7

[37] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024. 2, 3, 4, 6

[38] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024. 3

[39] Andriy Marusyk, Vanessa Almendro, and Kornelia Polyak. Intra-tumour heterogeneity: a looking glass for cancer? *Nature reviews cancer*, 12(5):323–334, 2012. 1, 4

[40] Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. In *International Conference on Learning Representations*, 2018. 3

[41] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 6

[42] Christopher Z Mooney, Robert D Duval, and Robert Duvall. *Bootstrapping: A nonparametric approach to statistical inference*. Number 95. sage, 1993. 1

[43] Xiaohuan Pei, Tao Huang, and Chang Xu. Efficientvmamba: Atrous selective scan for light weight visual mamba. *arXiv preprint arXiv:2403.09977*, 2024. 3, 4

[44] Yansheng Qiu, Delin Chen, Hongdou Yao, Yongchao Xu, and Zheng Wang. Scratch each other's back: Incomplete multi-modal brain tumor segmentation via category aware group self-support learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21317–21326, 2023. 1, 2, 7

[45] Santiago Ramón y Cajal, Marta Sesé, Claudia Capdevila, Trond Aasen, Leticia De Mattos-Arruda, Salvador J Diaz-Cano, Javier Hernández-Losa, and Josep Castellví. Clinical implications of intratumor heterogeneity: challenges and opportunities. *Journal of Molecular Medicine*, 98:161–177, 2020. 1, 4

[46] Jiacheng Ruan and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*, 2024. 3

[47] Dong She, Yueyi Zhang, Zheyu Zhang, Hebei Li, Zihan Yan, and Xiaoyan Sun. Eoformer: Edge-oriented transformer for brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 333–343. Springer, 2023. 2

[48] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 3

[49] Jimmy TH Smith, Andrew Warrington, and Scott Linderman. Simplified state space layers for sequence modeling. In *International Conference on Learning Representations*, 2022. 2

[50] Chloe Wang, Oleksii Tsepa, Jun Ma, and Bo Wang. Graph-mamba: Towards long-range graph sequence modeling with selective state spaces. *arXiv preprint arXiv:2402.00789*, 2024. 3

[51] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15878–15887, 2023. 6, 7

[52] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *Medical Image Computing*

and Computer Assisted Intervention–MICCAI, pages 109–119. Springer, 2021. 2

[53] Yixin Wang, Yang Zhang, Yang Liu, Zihao Lin, Jiang Tian, Cheng Zhong, Zhongchao Shi, Jianping Fan, and Zhiqiang He. Acn: adversarial co-training network for brain tumor segmentation with missing modalities. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*, pages 410–420. Springer, 2021. 1, 2

[54] Shicai Wei, Chunbo Luo, and Yang Luo. Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20039–20049, 2023. 1, 2, 7

[55] Shicai Wei, Yang Luo, Yuji Wang, and Chunbo Luo. Robust multimodal learning via representation decoupling. In *European Conference on Computer Vision*, pages 38–54. Springer, 2024. 3

[56] Jing Nathan Yan, Jiatao Gu, and Alexander M Rush. Diffusion models without attention. *arXiv preprint arXiv:2311.18257*, 2023. 3

[57] Chenhongyi Yang, Zehui Chen, Miguel Espinosa, Linus Ericsson, Zhenyu Wang, Jiaming Liu, and Elliot J Crowley. Plainmamba: Improving non-hierarchical mamba in visual recognition. *arXiv preprint arXiv:2403.17695*, 2024. 2, 3, 4

[58] Heran Yang, Jian Sun, and Zongben Xu. Learning unified hyper-network for multi-modal mr image synthesis and tumor segmentation with missing modalities. *IEEE Transactions on Medical Imaging*, 2023. 1, 2

[59] Qiushi Yang, Xiaoqing Guo, Zhen Chen, Peter YM Woo, and Yixuan Yuan. D 2-net: Dual disentanglement network for brain tumor segmentation with missing modalities. *IEEE Transactions on Medical Imaging*, 41(10):2953–2964, 2022. 2

[60] Chenyu You, Ruihan Zhao, Fenglin Liu, Siyuan Dong, Sandeep Chinchali, Ufuk Topcu, Lawrence Staib, and James Duncan. Class-aware adversarial transformers for medical image segmentation. pages 29582–29596, 2022. 2

[61] Ziqi Yu, Xiaoyang Han, Shengjie Zhang, Jianfeng Feng, Tingying Peng, and Xiao-Yong Zhang. Mousegan++: Unsupervised disentanglement and contrastive representation for multiple mri modalities synthesis and structural segmentation of mouse brain. *IEEE Transactions on Medical Imaging*, 2022. 1, 2

[62] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 107–117. Springer, 2022. 1, 2, 6, 7

[63] Zheyu Zhang, Xinzhao Liu, Zheng Chen, Yueyi Zhang, Huanjing Yue, Yunwei Ou, and Xiaoyan Sun. Anatomical consistency distillation and inconsistency synthesis for brain tumor segmentation with missing modalities. *arXiv preprint arXiv:2408.13733*, 2024.

[64] Zheyu Zhang, Gang Yang, Yueyi Zhang, Huanjing Yue, Aiping Liu, Yunwei Ou, Jian Gong, and Xiaoyan Sun. Tmformer: Token merging transformer for brain tumor segmentation with missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7414–7422, 2024. 2

[65] Sijie Zhao, Hao Chen, Xueliang Zhang, Pengfeng Xiao, Lei Bai, and Wanli Ouyang. Rs-mamba for large remote sensing image dense prediction. *arXiv preprint arXiv:2404.02668*, 2024. 2, 3

[66] Zechen Zhao, Heran Yang, and Jian Sun. Modality-adaptive feature interaction for brain tumor segmentation with missing modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 183–192. Springer, 2022. 1, 7, 2

[67] Zhuoran Zheng and Jun Zhang. Fd-vision mamba for endoscopic exposure correction. *arXiv preprint arXiv:2402.06378*, 2024. 3

[68] Chenhong Zhou, Changxing Ding, Xinchao Wang, Zhentai Lu, and Dacheng Tao. One-pass multi-task networks with cross-task guided attention for brain tumor segmentation. *IEEE Transactions on Image Processing*, 29:4516–4529, 2020. 2

[69] Tao Zhou, Huazhu Fu, Geng Chen, Jianbing Shen, and Ling Shao. Hi-net: hybrid-fusion network for multi-modal mr image synthesis. *IEEE Transactions on Medical Imaging*, 39 (9):2772–2781, 2020. 1, 2

[70] Tongxue Zhou, Stéphane Canu, Pierre Vera, and Su Ruan. Latent correlation representation learning for brain tumor segmentation with missing mri modalities. *IEEE Transactions on Image Processing*, 30:4263–4274, 2021. 1, 2

[71] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, 2019. 2

[72] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. 2024. 2, 3, 4, 6