Enhancing Modality-Agnostic Representations via Meta-Learning for Brain Tumor Segmentation — Supplementary Material —

Aishik Konwer[1], Xiaoling Hu[1], Joseph Bae[2], Xuan Xu[1], Chao Chen[2], Prateek Prasanna[2]
[1]Department of Computer Science, Stony Brook University
[2]Department of Biomedical Informatics, Stony Brook University
{akonwer,xiaolhu,xuaxu}@cs.stonybrook.edu
{joseph.bae,chao.chen.1,prateek.prasanna}@stonybrook.edu

In the supplementary material, we provide additional information to better understand the contributions and claims of our proposed work. First, the ablation results for various encoder-decoder backbones (3DUnet, nnUnet, AttentionUnet) are shown in Sec. 7. Our method also attains state-of-the-art performance when re-implemented with a 3DUnet backbone (like others). Note that we always maintain 50% full modality as the default setup for our approach. Other methods are however re-implemented with two different proportions (100% and 50% full modality data). In Sec. 8 we demonstrate the robustness of our approach to varying proportions of full modality data. Ablation results are provided for additional tumor regions. Evaluations via an additional metric, Hausdorff Distance, on BRATS2018 and BRATS2020 are shown in Sec. 9. Comprehensive results on BRATS2019 and BRATS2020 datasets are provided in Sec. 10. In Sec. 11, further experiments are conducted to test the bias of our model to the occurrence of a specific modality (FLAIR or T1c) as input during training. In Sec. 12 we discuss the fusion strategy in more detail through equations. Architectural details and experiments with different fusion baselines are demonstrated in Sec. 13. Additional qualitative segmentation maps are shown in Sec. 14. Further details regarding the implementation, including pre-processing steps, are outlined in Section 15. The segmentation performance of our model on additional non-BRATS datasets can be found in Sec. 16.

7. Ablation results on robustness to encoder-decoder backbones

Our proposed meta-learning and adversarial training strategies are independent of the backbones utilized in the framework. We evaluate our approach using different backbones including 3DUnet [2], nnUnet [6], and AttentionUnet [8]. The average DSCs reported in Tab. 7 vary marginally between 1.25% and 2.6% across all encoder-decoder variants, highlighting the backbone-agnostic nature of our framework. A schematic of the adopted Swin-UNETR encoder is provided in Fig. 11.

Methods	Average DSC(%), p-value (10^{-2})		
	WT	TC	ET
3DUnet [2]	85.70, 42.04	77.87, 67.28	59.93, 67.41
AttentionUnet [8]	86.02, 52.05	78.05, 71.10	60.46, 73.13
nnUnet [6]	86.53, 72.47	78.64, 86.16	62.28, 96.69
Ours	**87.12**	**79.12**	**62.53**

Table 7: Ablation on backbone variants

For the convenience of comparison, we have listed all the model performances (implemented using 3D-Unet backbone) in Tab 8. It should be noted that even with 3D-Unet as the backbone, our proposed method achieves results comparable to SOTA. This performance improvement may be attributed to the proposed meta and adversarial learning techniques, rather than the choice of backbone. However, our framework is trained with only 50% full modality samples, unlike other approaches that utilize full modality for all patients (100%).

Moreover, methods like mmFormer, RFNet, and ACN *always require full-modality data* as input. For a fair comparison, we have demonstrated in Tab. 9 that, if considering only 50% full-modality data as input (like ours), there is a significant drop in performance for all other methods.

8. Additional ablation results on robustness to full modality

Ablation results on the WT region have been provided in the main paper (Sec. 4.2, Fig. 7a) to demonstrate that our method performs well even with a limited number of full

Methods	Average DSC (%)		
	WT	TC	ET
HeMIS [5]	76.23	58.57	45.11
U-HVED [4]	79.55	64.55	48.44
D2-Net [10]	77.04	67.65	45.22
ACN [9]	85.25	77.16	**60.85**
RFNet [3]	85.75	76.99	59.67
mmFormer [11]	**86.25**	74.89	55.88
Ours (3D-Unet)	85.70	**77.87**	59.93

Table 8: Comparison on BRATS2018 with 3D-Unet backbone. All methods here are implemented with 3D-Unet. Only our approach is trained with 50% full modality samples while HeMIS, U-HVED. D2-Net, ACN, RFNet, and mmFormer are trained with 100% full modality samples.

Methods	Average DSC (%), p-value (10^{-2})		
	WT	TC	ET
ACN[9]	65.81, 0.01*	57.66, 0.01*	47.36, 1.50*
RFNet[3]	73.96, 0.01*	62.37, 0.01*	50.24, 3.95*
mmFormer[11]	75.34, 0.01*	66.19, 0.01*	52.81, 8.79
Ours (3D-Unet)	**85.70**	**77.87**	**59.93**

Table 9: Comparison (DSC%, p-value) on BRATS2018 with 3D-Unet backbone. All methods here are implemented with 3D-Unet, and trained with 50% full modality samples.

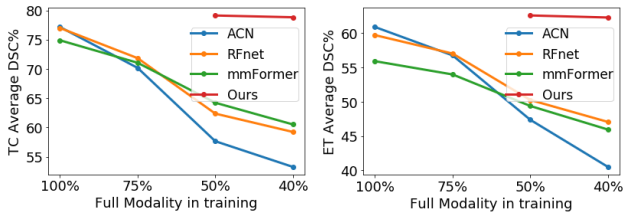

Figure 8: Ablation studies for varying % of full modality in training.

modality samples in training. Here we are providing additional results for TC and ET regions. We compare against ACN [9], RFNet [3], and mmFormer [11] by varying the full modality count from 100% to 40% (Fig. 8). In order to retain sufficient samples for each combination task in meta-training, we assume that at least 50% of the patients have partial modalities. Hence we show our results only on 50% and 40% proportions of full modality data. Unlike other methods, ours shows only a minor decline in DSC (0.3%) for both TC and ET. These experimental results further support the claim that our proposed method is robust to full modality setting.

We also present the results (Tab. 10) achieved by our method when trained with 10% and 20% full modality samples. Notably, our method still generates high dice scores even in such severely missing modality scenarios. Please note that it was not possible to train SOTA methods in this scenario since they are left with only ≈20 or ≈40 subjects.

Settings	Average DSC (%)		
	WT	TC	ET
10% FM	81.56	72.89	56.70
20% FM	84.41	76.63	60.82
50% FM	87.12	79.12	62.53

Table 10: Ablation on BRATS2018 when trained with an extremely low proportion of full modality samples.

9. Additional metric (Hausdorff distance)

Model evaluations have also been performed using Hausdorff Distance (HD95) on BRATS2018 and BRATS2020, respectively. The results can be found in Tab. 11 and 12. It can be observed from Tab. 11 that our method significantly outperforms SOTA in 2/3 tumor regions (WT, TC) and emerges second-best for ET on BRATS2018; noting that all other methods are trained with 100% full modality samples, ours is only 50%.

Methods	Average HD95 (↓), p-value (10^{-2})		
	WT	TC	ET
HeMIS [5]	14.85±7.32, 0.08*	15.58±8.44, 0.16*	19.65±12.37, 0.25*
U-HVED [4]	13.64±6.27, 0.12*	14.91±7.19, 0.09*	18.43±11.68, 0.42*
D2-Net [10]	10.82±6.70, 7.75	11.76±7.35, 5.12	14.79±8.79, 1.75*
ACN [9]	8.15±2.03, 39.05	9.37±2.81, 8.39	**8.62±2.43**, 86.77
RFNet [3]	7.89±1.72, 58.64	8.43±2.52, 42.09	12.56±3.68, 0.36*
mmFormer [11]	<u>7.67±2.14</u>, 84.97	<u>8.06±2.41</u>, 69.59	10.54±3.13, 11.41
Ours	**7.53±1.86**	**7.73±2.16**	<u>8.78±2.77</u>

Table 11: Comparison on BRATS2018 with HD95. The best and second best scores are **bolded** and <u>underlined</u>, respectively.

Methods	Average HD95 (↓), p-value (10^{-2})		
	WT	TC	ET
HeMIS [5]	14.41±7.14, 0.11*	15.13±8.29, 0.21*	19.24±12.07, 0.26*
U-HVED [4]	13.32±6.11, 0.13*	14.74±6.97, 0.08*	18.26±11.53, 0.41*
RFNet [3]	<u>7.66±1.74</u>, 76.06	<u>8.27±2.45</u>, 44.00	<u>12.38±3.72</u>, 0.44*
Ours	**7.46±1.82**	**7.60±2.23**	**8.69±2.72**

Table 12: Comparison on BRATS2020 with HD95. The best and second best scores are **bolded** and <u>underlined</u>, respectively.

10. Results on BRATS2019 and BRATS2020 datasets

In Tab. 13, we compare our approach with three state-of-the-art methods including HeMIS [5], U-HVED [4], and RFNet [3] for tumor segmentation on BRATS2020 dataset [7]. The average DSCs of the three tumor areas are boosted by 1.78%, 2.84%, and 3.13%, respectively. A similar comparison on the BRATS2019 dataset is shown in Tab. 14, where the average DSC scores are boosted by 1.57%, 2.83%, and 2.94%.

M																		Avg
	FLAIR	○	○	○	●	○	○	●	○	●	●	●	●	●	○	●		
	T1	○	○	●	○	○	●	●	●	○	○	●	●	○	●	●		
	T1c	○	●	○	○	●	●	○	○	○	●	●	○	●	●	●		
	T2	●	○	○	○	●	○	○	●	●	○	○	●	●	●	●		
WT	HeMIS[5]	80.34	66.92	66.35	58.72	85.16	73.41	69.79	83.30	83.76	73.41	76.78	84.43	85.17	85.84	86.03		77.29
	U-HVED[4]	82.13	71.42	58.30	82.76	85.72	74.09	86.46	84.34	87.91	87.15	86.59	88.66	88.92	85.86	89.43		82.65
	RFNet[3]	86.30	76.34	77.72	87.05	88.02	81.07	89.72	88.02	89.64	89.51	90.44	90.62	90.55	88.50	91.01		<u>86.96</u>
	Ours	**88.24**	**82.29**	**83.41**	**88.37**	**88.78**	**83.26**	**90.52**	**89.66**	**90.55**	**90.83**	**91.34**	**91.68**	**91.17**	**89.49**	**91.57**		**88.74**
TC	HeMIS[5]	60.83	74.22	48.57	37.03	79.84	78.35	48.19	60.80	60.21	74.62	78.88	63.48	79.24	81.56	81.03		67.12
	U-HVED[4]	61.37	74.93	39.54	52.42	80.27	79.11	57.38	62.17	63.47	77.45	79.02	65.39	80.19	81.72	81.68		69.07
	RFNet[3]	70.94	82.45	65.58	69.88	85.82	83.88	72.76	72.90	73.45	85.71	85.97	74.74	86.11	85.55	86.24		<u>78.79</u>
	Ours	**73.56**	**86.37**	**74.69**	**72.33**	**87.71**	**87.52**	**75.94**	**74.50**	**76.24**	**87.79**	**87.82**	**76.93**	**87.31**	**87.98**	**87.75**		**81.63**
ET	HeMIS[5]	32.78	64.95	20.41	14.63	71.12	71.40	19.04	29.76	30.66	69.52	71.39	32.13	71.98	72.37	72.44		49.64
	U-HVED[4]	31.86	68.43	18.21	25.85	70.48	70.79	27.94	32.37	33.64	71.24	72.16	34.48	71.72	71.92	71.87		51.53
	RFNet[3]	48.03	74.84	36.58	38.45	76.66	76.52	43.12	51.40	51.02	76.38	77.10	49.82	77.07	78.10	77.02		<u>62.14</u>
	Ours	**52.77**	**80.06**	**42.28**	**44.87**	**78.92**	**79.85**	**46.73**	**54.67**	**54.29**	**78.81**	**77.31**	**50.69**	**79.24**	**79.43**	**79.12**		**65.27**

Table 13: Comparison with state-of-the-art for the different combinations of available modalities on BRATS2020. Dice scores (DSC %) are computed for three nested tumor subregions - Whole tumor (WT), Tumor core (TC), Enhancing tumor (ET). Modalities present are denoted by ●, the missing ones by ○. The best and second best scores are **bolded** and <u>underlined</u>, respectively.

Methods	Average DSC (%)		
	WT	TC	ET
HeMIS [5]	76.69	64.37	48.24
U-HVED [4]	81.53	67.81	50.25
RFNet [3]	86.49	77.92	60.88
Ours	**88.06**	**80.75**	**63.82**

Table 14: Comparison (DSC %) on BRATS2019

11. Additional ablation results on bias to presence of a specific modality

In the main paper (Sec. 4.2, Fig. 7b) we have provided ablation results on the WT region by varying FLAIR proportion in training from 35% to 45%. Here we provide extensive results for the remaining two tumor regions (TC and ET). Fig. 9 suggests that upon increasing FLAIR from 35% to 45%, our model's DSC gain (for both TC and ET) is much less when compared to that of U-HVED or D2Net. This demonstrates that our approach is not sensitive to presence of any particular modality. Similar conclusions can also be drawn when experiments are carried out keeping T1c as the rarely occurring modality instead of FLAIR. The results are presented in Tab. 15 and Fig. 10.

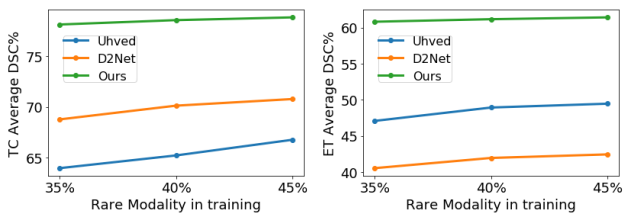

Figure 9: Ablation studies for varying % of FLAIR in training.

M										Avg
	FLAIR	○	○	○	●	●	●	○	●	
	T1	○	○	●	○	●	○	●	●	
	T1c	●	●	●	●	●	●	●	●	
	T2	○	●	○	○	○	●	●	●	
WT	U-HVED	54.38	77.63	63.70	82.28	84.06	84.96	81.19	86.37	76.82
	D2-Net	39.14	81.57	62.77	86.32	85.98	87.85	82.40	87.53	76.70
	Ours	**78.46**	**85.71**	**78.83**	**89.22**	**89.64**	**90.08**	**87.19**	**90.57**	**86.21**
TC	U-HVED	61.16	70.87	62.21	70.65	72.24	70.59	74.52	70.87	69.13
	D2-Net	61.73	78.46	75.31	80.69	78.17	79.85	78.77	79.58	76.57
	Ours	**83.62**	**84.79**	**83.56**	**84.18**	**84.35**	**83.88**	**85.80**	**85.94**	**84.51**
ET	U-HVED	57.23	65.39	62.08	65.76	68.15	67.81	67.26	69.42	65.38
	D2-Net	65.13	66.37	67.84	65.41	65.06	65.33	67.98	66.27	66.17
	Ours	**77.02**	**75.98**	**76.43**	**76.25**	**75.11**	**76.67**	**75.89**	**77.54**	**76.36**

Table 15: Ablation results for rare occurrence (35%) of T1c in training. ● for T1c in all combinations denote that T1c is always present in inference despite being rare on training.

12. Details on Feature Aggregation Module

For a particular level l, a modality feature $\mathbf{F}_j^l \in \mathbb{R}^{C\times H\times W\times Q}$ includes C channels and feature maps of size $H \times W \times Q$ where $j \in \{1, 2, ..., n\}$. The channels in these generated features are considered to encode relevant tumor-class specific information. Our fusion block exploits the correlation among available modality representations to develop a unified feature that best describes the tumor characteristics of a particular patient. First, the channel information γ_j^l of a modality at level l is preserved through the following equation:

$$\gamma_j^l = GAP(\mathbf{F}_j^l) = \frac{1}{H \times W \times Q} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{q=1}^{Q} \mathbf{F_j^l}(h, w, q), \tag{1}$$

where $j \in \{1, 2, ..., n\}$ and GAP denotes Global Average Pooling operation. Following this, we not only concatenate

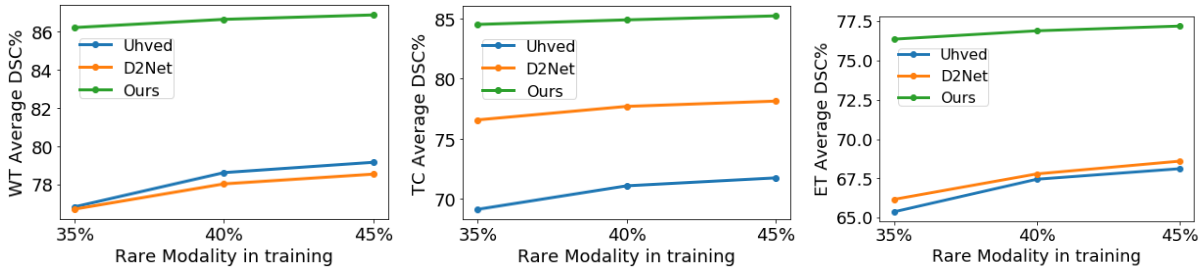

Figure 10: Ablation studies for varying % of T1c in training.

$\gamma_1^l, \gamma_2^l, ..., \gamma_n^l$, but also impute zeros in the channel information of $(M - n)$ missing modalities to form a resultant M-dimensional vector γ^l.

$$\gamma^l = \gamma_1^l \oplus \gamma_2^l \oplus \gamma_3^l ... \oplus \gamma_M^l, \tag{2}$$

γ^l is mapped to the channel weights of M modality features through a multi-layer perceptron (MLP) and sigmoid activation function, σ.

$$\Gamma^l = \sigma(MLP(\gamma^l)). \tag{3}$$

Though Γ^l contains M scalar values, only the weights of n available modalities are multiplied with their corresponding features. These weighted features are finally summed to obtain the fused representation $\mathbf{F}_{fused}^l$.

$$\mathbf{F}_{fused}^l = \sum_{j=1}^{n} \Gamma_j^l \mathbf{F}_j^l. \tag{4}$$

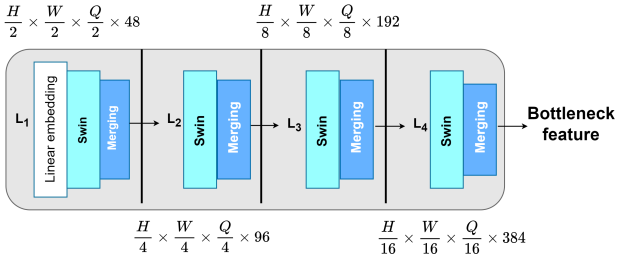

Figure 11: A schematic of the adopted Swin-UNETR encoder.

13. Fusion baselines and ablation

We design three baseline aggregation modules to highlight the contribution of our fusion strategy. The architectures of the three fusion baselines, (a) Sum, (b) Average, and (c) Att-Pool are illustrated in Fig. 12. For the first two approaches, feature maps from the available modalities are summed or averaged along the channel dimension C to obtain the fused feature. In the third approach, available modality features are individually passed through a Global Average Pooling (GAP) layer. The GAP outputs are fed to a Fully Connected Network (FCN) followed by a softmax activation function, producing the attention weights of each modality. Finally, attention-weighted summation of the original modality features gives rise to the fused feature. The ablation results are shown in Tab. 16. Our feature aggregation block provides a better technique for dynamically learning from the heterogeneous input modalities, followed by inducing channel interaction among them. However, this plug-and-play fusion module is not a primary contribution and can be replaced by SOTA fusion techniques [3, 1].

Methods	Average DSC (%)		
	WT	TC	ET
Sum	85.99	78.21	60.85
Average	86.14	78.36	61.30
Att-pool	86.93	79.07	62.28
Ours	**87.12**	**79.12**	**62.53**

Table 16: Ablation study on fusion.

14. Qualitative comparison

In Fig. 13 we visualize the segmentation masks predicted by U-HVED, RFNet, and our method from four combinations of modalities in the inference phase. Unlike other methods, our segmentations do not degrade sharply as additional modalities are dropped during inference. Even with single T2 or T1+T2 modalities, our model achieves higher DSC scores.

15. Additional pre-processing and implementation details

As part of pre-processing, the organizers skull-stripped the volumes and interpolated them to an isotropic 1mm^3 resolution. For a given patient, the four sequences have been co-registered to the same anatomical template. Augmentations including random rotations, intensity shifts, and mirror flipping, are applied to the resized images. The foreground voxels within the brain are intensity-normalized to

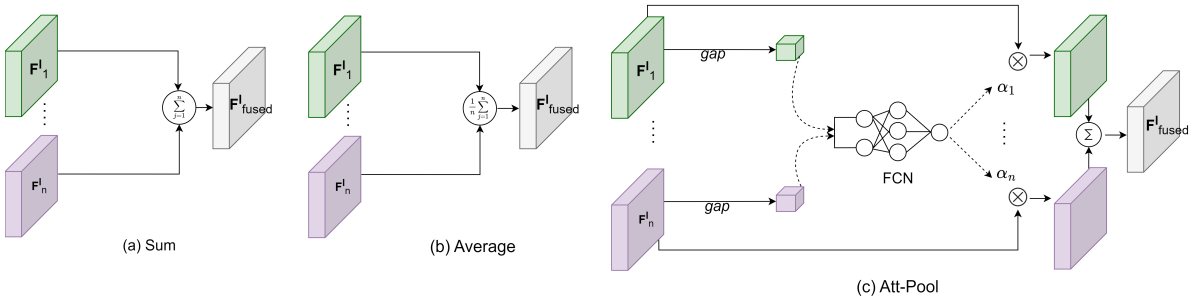

Figure 12: Fusion baselines

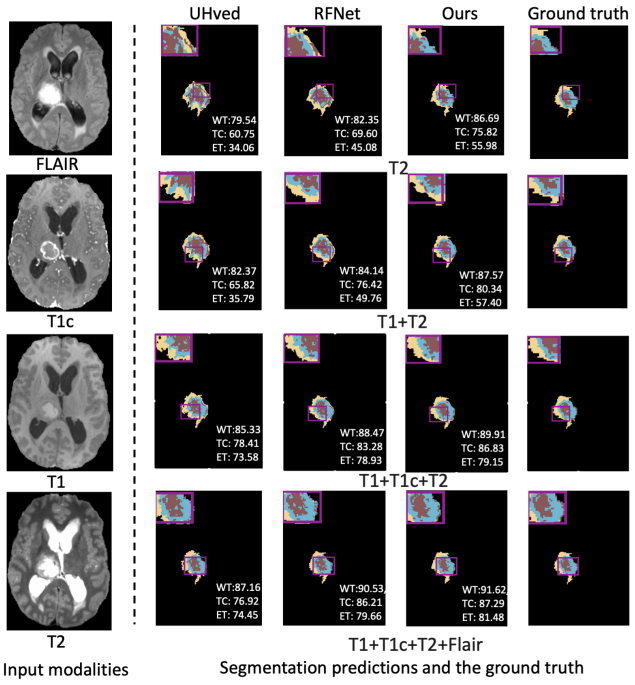

Figure 13: Qualitative comparisons with SOTA. Column 1: four MRI modalities. Column 2-4: segmentation maps predicted by three methods for different combinations of modalities. Column 5: Ground truth.

zero mean and unit standard deviation. We train our network using AdamW optimizer with an outer loop learning rate $\beta = 5e-4$ for a maximum of 500 epochs. The two hyperparameters λ_1 and λ_2 in generator loss $\mathcal{L}_E$ are taken as 0.8 and 0.2, respectively. During training, $\mathcal{L}_{dis}$ is multiplied by 0.5 to prevent it from overpowering the generator.

Learning rate (LR: 5e-5), λ_1: 0.8, λ_2: 0.2, and scale of discriminator (Sc: 0.5) were selected based on the model performance. Results with different sets of parameters are shown in Tab. 17.

LR	WT DSC(%)	λ_1, λ_2	Avg DSC % (WT, TC, ET)	Sc	WT DSC(%)
5e-3	86.79	0.9, 0.1	86.89, 78.94, 62.37	0.25	86.44
5e-4	86.95	0.8, 0.2	87.12, 79.12, 62.53	0.5	87.12
5e-5	87.12	0.7, 0.3	86.97, 78.83, 62.19	0.6	87.03

Table 17: Selection of experimental parameters

16. Results on additional datasets

We show the segmentation results (Tab. 18, 19) on two additional datasets not in the BRATS cohort. The first dataset, D_1 contains 4 MRI modalities for 80 patients. For this dataset, we segment brain glioma tumors into 3 regions (WT, TC, ET). Another dataset, D_2 contains 1 MRI modality (FLAIR) and 1 CT modality for 85 patients with metastatic brain tumors as the segmentation targets. Unlike the solitary brain tumors studied in the other datasets, multiple distinct metastatic targets can occur at multiple locations within a patient's brain for D_2.

Methods	Average DSC (%)		
	WT	TC	ET
U-HVED [14]	75.37	60.29	47.52
RFnet [13]	81.04	72.15	53.22
mmFormer [55]	81.73	71.31	51.49
Ours	**82.53**	**74.26**	**56.13**

Table 18: Results on D_1.

FLAIR ● ○ ● ● / CT ○ ● ●	Avg DSC
U-HVED	49.93	46.20	48.59	48.24
RFNet	53.62	51.37	53.16	52.71
mmFormer	54.88	52.85	54.63	54.12
Ours	**55.19**	**53.27**	**55.06**	**54.50**

Table 19: Results on D_2.

References

[1] 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 *MICCAI*, 2019. 4

[2] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *MICCAI*, 2016. 1

[3] Yuhang Ding, Xin Yu, and Yi Yang. RFNet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In *ICCV*, 2021. 2, 3, 4

[4] Reuben Dorent, Samuel Joutard, Marc Modat, Sébastien Ourselin, and Tom Vercauteren. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In *MICCAI*, 2019. 2, 3

[5] Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. HeMIS: Hetero-modal image segmentation. In *MICCAI*, 2016. 2, 3

[6] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 2021. 1

[7] 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). *TMI*, 2014. 2

[8] Ozan Oktay, Jo Schlemper, Loic Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention U-Net: Learning where to look for the pancreas. In *MIDL*, 2018. 1

[9] 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 *MICCAI*, 2021. 2

[10] Qiushi Yang, Xiaoqing Guo, Zhen Chen, Peter YM Woo, and Yixuan Yuan. D2-net: Dual disentanglement network for brain tumor segmentation with missing modalities. *TMI*, 2022. 2

[11] 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 *MICCAI*, 2022. 2