Query Re-Training for Modality-Gnostic Incomplete Multi-modal Brain Tumor Segmentation

Delin Chen$^{1,2[0000-0002-9519-093X]}$, Yansheng Qiu$^{1,2[0000-0002-5619-0902]}$, and Zheng Wang$^{1,2\star[0000-0003-3846-9157]}$

[1] National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, School of Computer Science, Wuhan University
[2] Hubei Key Laboratory of Multimedia and Network Communication Engineering
{chendelin, qiuyansheng, wangzwhu}@whu.edu.cn

Abstract. Although Magnetic Resonance Imaging (MRI) is crucial for segmenting brain tumors, it frequently lacks specific modalities in clinical practice, which limits prediction performance. In current methods, training involves multiple stages, and encoders are different for each modality, which means hybrid modules must be manually designed to incorporate multiple modalities' features, lacking interaction across modalities. To ameliorate this problem, we propose a transformer-based end-to-end model with just one auto-encoder to provide interactive computations in any modality missing condition. Considering that it is challenging for a single model to perceive several missing states, we introduce learnable modality combination queries to assist the transformer decoder in adjusting to the incomplete multi-modal segmentation. Furthermore, to address the suboptimization issue of the Transformer under small datasets, we adopt a re-training mechanism to facilitate convergence to a better local minimum. The extensive experiments on the BraTS2018 and BraTS2020 datasets demonstrate that our method outperforms the current state-of-the-art methods for incomplete multi-modal brain tumor segmentation on average.

Keywords: Query · Re-Training · Incomplete Multi-modal · Brain Tumor Segmentation.

1 Introduction

Magnetic resonance image (MRI) segmentation plays an integral role in quantitative brain tumor image analysis, which is designed for different tissues of brain structures and brain tumors with multiple imaging modalities, such as Fluid Attenuation Inversion Recovery (FLAIR), contrast enhanced T1-weighted (T1c), T1-weighted (T1) and T2-weighted (T2). It has been demonstrated that simultaneously combining four modalities could improve multi-modal MRI performance for brain tumor segmentation [25,13,17,23,8]. Nevertheless, missing

[⋆] Corresponding Author

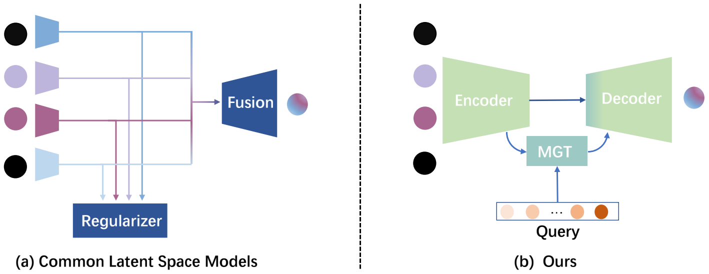


Fig. 1. Incomplete Multi-modal Brain Tumor Segmentation Frameworks (colored for residual, ● for missing) - (a) Common Latent Space Models, *i.e.*, RFNet [5] proposes a framework with separate encoders for each modality, a decoder and hierarchical fusion blocks. The four encoders are arranged in a top-to-bottom sequence corresponding to the four different modalities: Flair, T1c, T1, and T2. (b) Our proposed method has a single encoder, a decoder and Modality-Gnostic Transformer (MGT) Modules that learn modality combination queries to solve all conditions effectively.

modalities are prevalent in clinical practice due to data corruption, different scanning protocols, and patient conditions [21,11,15,14] , severely reducing previous segmentation algorithms' performance. Therefore, a robust multi-modal approach is required for flexible, practical clinical applications to address the issue of missing one or more modalities.

Current research on incomplete medical image segmentation [5,3,22,10] primarily considers improving the network's ability to extract features from different modalities by separate encoders and producing discriminative fusion features for segmentation (Figure 1 (a)). Nevertheless, only individual features are learned, and the information across modalities cannot interact, increasing the difficulty of feature fusion [24]. To tackle this problem, we introduce a unified transformer-based encoder that allows direct interaction between different modalities. This approach entails concatenating different modalities and feeding them into the encoder, whereby the design of self-attention allows for the natural interaction of the input. However, unified architectures make it difficult to perceive multiple modalities scenarios and degrade the performance. Fortunately, by exploiting the properties of the attention mechanism, Valanarasu *et al.* [18] propose weather-type learnable embeddings to tackle all adverse weather removal problems in a single encoder-decoder transformer network. Deriving from the random initialization of the learnable embeddings in [18], we codify the modally missing combinations and initialize the learnable embeddings with them, which can provide more informative guidance.

Furthermore, Vision Transformers need a lot of data for training, usually more than what is necessary to standard CNNs [7]. The transformer-based models started with randomly initialized parameters, may easily over-fit a small number of training pairs and make the model be trapped into a poor local minimum. Inspired by [16], adopt a re-training mechanism to facilitate convergence to a better local minimum.

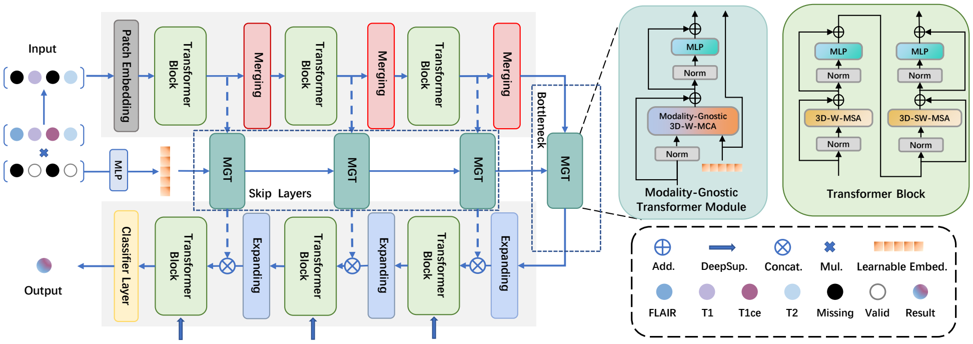


Fig. 2. Illustration of our proposed QuMo architecture. QuMo comprises three primary modules: A transformer encoder to extract hierarchical features, Modality-Gnostic transformer modules and a transformer decoder. A 3D volume concatenated by four different modality volumes multiplied by a modal code is applied to simulate different modalities missing states.

To this end, we propose **QuMo**: **Qu**ery re-training for **Mo**dality-gnostic incomplete multi-modal brain tumor segmentation to tackle all modality-missing states simultaneously (Figure 1 (b)). Specifically, our contributions are threefold:

- We propose **QuMo** : **Qu**ery re-training for **Mo**dality-gnostic incomplete multi-modal brain tumor segmentation, an effective solution with only one encoder and one decoder which can provide the direct interaction of different modalities within the network.
- We propose a Modality-Gnostic transformer module with learnable modality combination embeddings as queries to effectively handle all the modality-missing states, making the decoder be aware of different modality combinations. Furthermore, we adopt a query re-training mechanism to facilitate the model convergence to a better local minimum under small datasets.
- Taking advantage of the proposed method, **QuMo** could achieve state-of-the-art performance on the used BraTs2018 and BraTs2020 benchmarks.

2 Method

2.1 Architecture Overview

An overview of QuMo is illustrated in Figure 2, QuMo contains a transformer-based unified architecture to accept all valid modalities as input simultaneously, which can provide the direct interaction of different modes within the encoder. Following previous methods [1,12,13], we exploit the transformer [19] architecture for explicitly long-range contextual modeling within the input MRI modalities effectively In Vision Transformer (ViT) [7], tokens are required to contain spatial information due to the way they are constructed and the significance of performing self-attention by windowing in ViT has been demonstrated in several recent studies, most notably in Swin Transformer [12]. In particular, the encoder consists of a patch embedding layer and patch merging layers followed by

Transformer blocks. The decoder is designed to generate the segmentation mask based on four output feature maps of different resolutions from Modality-Gnostic Transformer modules (MGTs). In implementations, the transformer blocks in the decoder follow the same design as the encoder, and we deviate from the encoder by patch expanding layers and convolutional classifier layers.

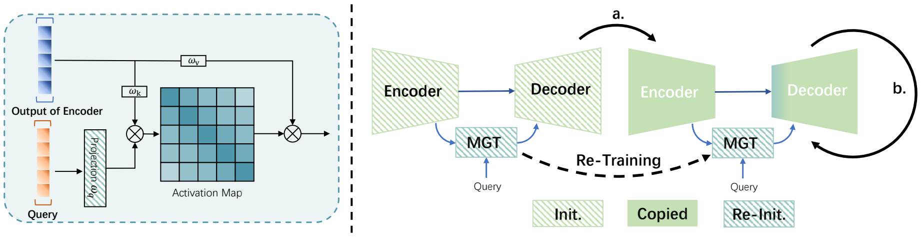


Fig. 3. *Left*: **Configuration of the Modality-Gnostic 3D-W-MCA in the MGTs.** The queries here are learnable embeddings representing the modality combination, while the keys and values are features taken from the output of the transformer encoder. *Right*: **The process of our re-training mechanism.** The parameters of the module with slash color backgrounds are initialized as the original rules, while those with solid background are copied from the former training phase. Noteworthy, only the projection layers in MGTs are re-Initialized.

2.2 Modality-Gnostic Transformer Module

An autoregressive decoder is used in the original transformer decoder [19] to predict the output sequence one element at a time. However, the model's inability to perceive the modal input state makes dynamic handling of different missing modalities difficult. Detection transformer (DETR) [2] uses object queries to decode the box coordinates and class labels to produce the final predictions. Moreover, TransWeather [18] uses weather type queries to decode different restore tasks, predict a task feature vector and use it to restore the clean image. Inspired by them, we define modality combination embeddings as queries to guide the decoder to perceive different modality-missing situations. The queries are learnable embeddings that attend to the feature outputs from the encoder and are learned along with other model parameters during the training phase, illustrated in Figure 2. Unlike the self-attention transformer block where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are taken from the same input, in Modality-Gnostic 3D window multi-head cross-attention (W-MCA), $\mathbf{Q}$ is a modality combination learnable embedding. At the same time, $\mathbf{K}$ and $\mathbf{V}$ are tokens from the features taken from the corresponding stage of the transformer encoder after linear mapping, illustrated in Figure 3. The computation in the MGT can be summarized as:

$$\hat{\mathbf{I}}^l = \text{3D-W-MCA}\left(\text{LN}\left(\mathbf{I}^{l-1}\right), \mathbf{Q}\right) + \mathbf{I}^{l-1}, \mathbf{I}^l = \text{MLP}(\text{LN}(\hat{\mathbf{I}}^l)) + \hat{\mathbf{I}}^l \qquad (1)$$

where $\mathbf{Q}$ denotes the learnable queries, LN refers to Layer Normalization, $\hat{\mathbf{I}}^l$ and $\mathbf{I}^l$ denote the output features of the MCA module and the MLP module for layer

l, respectively. The 3D-W-MCA is mathematically described as follows:

$$\mathbf{K},\mathbf{V}=\omega_{k,v}(\mathrm{LN}\left(\mathbf{I}^{l-1}\right)),\hat{\mathbf{Q}}=\omega_q(\mathbf{Q}),\hat{\mathbf{I}}^l=\mathrm{Softmax}(\frac{\hat{\mathbf{Q}}\mathbf{K}^\top}{\sqrt{d}})\mathbf{V}+\mathbf{I}^{l-1} \tag{2}$$

where $\omega_{k,v},\omega_q$ are projection functions to produce $\mathbf{K}$, $\mathbf{V}$ and $\hat{\mathbf{Q}}$ respectively, d represents the number of the tokens' channels.

In this paper, the proposed method contains four Modality-Gnostic transformer modules, among which three are at skip-connection layers, and one acts as the bottleneck layer. The output decoded features are fused with the features extracted across the MGTs through skip connections and the bottleneck layer at each stage. At the beginning of the training phase, the learnable embeddings are initialized with modality-code $C=\{C_1,\ldots,C_N\}$, where $C_i\in\{0,1\}$ represents whether the C_i modality is missing or not and N represents the number of modalities. Specifically, we map C through fully connected layers so that the extended query embeddings can perform matrix multiplication operations with $\mathbf{K}$.

2.3 Query Re-Training Strategy

Motivated by the observation that a segmentation model may converge to a better local minimum by equipping the Transformer encoder-decoder with better-initialized parameters [16], we design the query retraining mechanism. After the initial training, we continuously update the encoder and decoder parameters during training, while periodically resetting the MGTs' parameters, specifically the query projection, to encourage improved optimization.

As depicted in Figure 3, we first randomly initialize the entire model, denoted as process **a**, and terminate training when validation performance stabilizes. To further enhance model performance, we then proceed to continuously train the encoder and decoder after the first initial training while resetting the MGT modules to avoid convergence to sub-optimal local minima, denoted as process **b**. This process is repeated periodically until the best possible performance is achieved.

2.4 Loss Function

The segmentation results are learned under the supervision of the ground truth. Specifically, we supervise the transformer blocks in the decoder in a stage-wise manner. This deep supervision strategy [20] lets the transformer blocks focus on meaningful semantic regions at different scales. The training loss is based on the combination of a weighted cross-entropy loss $\mathcal{L}_{WCE}$ [3] to address the imbalance of different regions and a Dice loss $\mathcal{L}_{DL}$, expressed as:

$$\mathcal{L}=\sum_{i=1}^{N}\left(\mathcal{L}_{WCE}\left(\hat{y}_i,y_i\right)+\mathcal{L}_{DL}\left(\hat{y}_i,y_i\right)\right), \tag{3}$$

where N denotes the number of training data, $\hat{y}_i$ and y_i denote predicted segmentation results and the ground-truth. $\mathcal{L}_{WCE}$ and L_{DL} are formulated as:

$$\mathcal{L}_{WCE} = \sum_{k \in K} \frac{\left\| -\omega_k \cdot y_k \cdot \log\left(\hat{y}_k\right) \right\|_1}{H \cdot W \cdot Z}, \mathcal{L}_{DL} = 1 - \sum_{k \in K} \frac{2 \cdot \left\| \hat{y}_k \bigcap y_k \right\|_1}{K_{\text{num}} \ \cdot \left(\left\| \hat{y}_k \right\|_1 + \left\| y_k \right\|_1 \right)}, \tag{4}$$

where $\| \cdot \|_1$ denotes the L1 norm, and H, W, Z denote the height, width and slice number of the 3D volumes, respectively. K denotes the set of brain tumor regions, including BG (background), NCR/NET, ED and ET. ω_k is the weight for the region k and $\omega_k = 1 - \frac{\|y_k\|_1}{\sum_{k' \in K} \|y_{k'}\|_1}$. $\bigcap$ denotes the overlap between predictions and ground-truth masks, and K_{num} denotes the number of regions in K.

3 Experiments

3.1 Implementation Details

Datasets We evaluate our method on the Multi-modal Brain Tumor Segmentation Challenge 2018 (BraTS2018) dataset and the BraTs2020 dataset. Each subject in the dataset contains four MRI contrasts (FLAIR, T1c, T1, T2), following the challenge, there are three segmentation classes, including whole tumor ("complete"), core tumor ("core") and enhancing tumor ("enhancing"). The ground truth was obtained by expert manual annotation.

Experimental Setup For the image pre-processing, the MRI images are skull-stripped, co-registered and re-sampled to $1mm^3$ resolution by the data collector. In this work, following [3], we additionally cut out the black background area outside the brain and normalize each MRI modality to zero mean and unit variance in the brain area. During training, input images are randomly cropped to 128 × 128 × 128 and are then augmented with random rotations, intensity shifts and mirror flipping. We train our network with a batch size of 1 in three re-training cycles. Adam optimizer [9] with a cosine scheduler is leveraged to optimize the network with β_1 and β_2 of 0.9 and 0.999 respectively.

3.2 Performance Comparison

To evaluate the performance, we compare our model with four state-of-the-art methods using a commonly used performance metric *Dice* [4], including U-HVED [6], RobustSeg [3], RFNet [5] and mmformer [22], all experiments were conducted employing the same train and test split lists as U-HVED [6] on BRATS2018 and RFNet [5] on BRATS2020 for a fair comparison.

As shown in Table 1 and Figure 4, our method achieves superior segmentation performance. For example, compared with the previous state-of-the-art method, *i.e.*, RFNet [5], our QuMo improves the average Dice scores by 1.10%, 1.66% and 3.93% in the whole tumor, the tumor core and the enhancing tumor, respectively. Moreover, our method outperforms the state-of-the-art methods on the vast

Table 1. Results of state-of-the-art unified models (U-HVED [6],RobustSeg [3], RFNet [5],mmformer [22]) and our method QuMo, on BraTS 2018 dataset. Dice similarity coefficient (DSC) [%] is employed for evaluation with every combination settings of modalities. ● and ○ denote available and missing modalities, respectively. The results with underlined denote the second best and with **bold** shows the best performance.

Modalities				Dice(%)														
				Complete					Core					Enhancing				
F	T1	T1c	T2	[6]	[3]	[5]	[22]	Ours	[6]	[3]	[5]	[22]	Ours	[6]	[3]	[5]	[22]	Ours
○	○	○	●	80.90	82.24	85.10	83.90	**86.06**	54.10	57.49	66.95	66.20	**70.53**	30.80	28.97	**44.56**	38.81	42.15
○	○	●	○	62.40	73.31	73.61	74.77	**77.60**	66.70	76.83	80.29	79.92	**81.09**	65.50	67.07	68.10	72.28	**75.55**
○	●	○	○	52.40	70.11	74.79	74.24	**77.51**	37.20	47.90	65.23	62.26	**65.80**	13.70	17.29	34.02	31.34	**40.65**
●	○	○	○	82.10	85.69	85.79	86.00	**89.25**	50.40	53.57	62.57	60.82	**68.28**	24.80	25.69	35.29	33.47	**44.28**
○	○	●	●	82.70	85.19	85.62	85.48	**86.75**	73.70	80.20	82.35	82.46	**83.63**	70.20	69.71	72.53	73.64	**74.97**
○	●	●	○	66.80	77.18	77.53	78.35	**79.36**	69.70	78.72	81.34	81.82	**82.04**	67.00	69.06	73.72	74.81	**76.00**
●	●	○	○	84.30	88.24	88.99	88.26	**90.10**	55.30	60.68	72.22	68.67	**74.19**	24.20	32.13	43.29	35.96	**49.02**
○	●	○	●	82.20	84.78	85.37	85.35	**86.59**	57.20	62.19	71.07	68.51	**73.18**	30.70	32.01	46.06	40.83	**46.37**
●	○	○	●	87.50	88.28	89.28	88.72	**90.37**	59.70	61.16	71.75	67.90	**74.22**	34.60	33.84	47.07	40.20	**48.56**
●	○	●	○	85.50	88.51	**89.39**	88.61	89.32	72.90	80.62	81.56	81.66	**83.28**	70.30	70.30	73.50	74.09	**77.34**
●	●	●	○	86.20	88.73	**89.87**	88.54	89.24	74.20	81.06	82.27	82.63	**83.64**	71.10	70.78	72.78	74.45	**77.46**
●	●	○	●	88.00	88.81	90.00	89.20	**90.51**	61.50	64.38	74.02	70.24	**74.76**	34.10	36.41	45.75	39.67	**52.56**
●	○	●	●	88.60	89.27	**90.36**	89.39	89.69	75.60	80.72	82.56	82.41	**83.41**	71.20	70.88	74.14	74.08	**77.08**
○	●	●	●	83.30	86.01	86.13	85.78	**86.91**	75.30	80.33	82.87	80.33	**83.33**	71.10	70.10	72.84	71.10	**76.53**
●	●	●	●	88.80	89.45	**90.59**	89.45	89.66	76.40	80.86	82.94	80.86	**83.58**	71.70	71.13	72.90	71.70	**77.05**
Average				80.10	84.39	85.49	85.07	**86.59**	64.00	69.78	76.00	74.75	**77.66**	50.00	51.02	58.44	56.95	**62.37**

majority of fifteen multi-modal combinations, including 11 out of 15 cases for the whole tumor, all cases for the core tumor, 14 out of 15 cases for the enhancing tumor. The quantitative results show that our QuMo brings more significant growth for enhancing tumor region, which are more challenging to segment, particularly improve the Dice scores by 8.99% when only Flair modality exists. We undertake additional validation to verify the efficacy of our model on the Brats2020 dataset. The results illustrated in Table 2 show our method yields superior performance compared to the State-of-the-Art (SOTA).

We conduct a comparison of computational complexity and model size. The result in Table 6 shows that our method is smaller than other algorithms in FLOPs(G) and smaller than transformer-based algorithms mmFormer[22] in model parameters. Visualization results in Figure 6 illustrate that our method is able to segment brain tumors well in various missing scenarios. For example, QuMo predicts an accurate segmentation map with only the T2 modal image. As the number of modes increases, the performance of the model becomes progressively better, and the performance in some severely missing cases is close to that in the full mode, e.g.T2 and F+T1. These results demonstrate the superiority of our method for incomplete multimodal learning of brain tumor segmentation.

3.3 Ablation Study

In this section, we investigate the MGT module, deep supervision and the retraining strategy, which are the key components of our method. All ablation experiments were conducted on the BraTS2018 dataset. We first set up a baseline network (“ Baseline ”) that does not use any MGT modules or deep supervision

Table 2. Results of previous models and our method on BraTS 2020 dataset.

Methods	Dice(%)		
	Comp.	Core	En.
U-HVED [6]	81.24	67.19	48.55
Robust [3]	84.17	73.45	55.49
RFNet [5]	86.98	78.23	61.47
mmFormer [22]	86.49	76.06	63.19
Ours	**87.65**	**78.37**	**63.21**

Table 3. Ablation study of critical components of QuMo.

MGT		D.S.	Init.	Average Dice(%)
Bottle.	Skip.			
✗	✗	✗	-	71.09
✗	✗	✓	-	73.27
✓	✗	✗	Rand	73.62
✓	✗	✓	Rand	75.01
✓	✓	✓	Rand	75.10
✓	✓	✓	Code	**75.54**

Table 4. The number of queries.

Number	Dice(%)			
	Complete	Core	Enhancing	Average
0	85.93	76.24	57.76	73.27
200	85.73	75.90	59.87	73.83
300	86.26	76.42	59.87	74.18
400	86.45	77.61	61.98	75.34
500	86.59	77.66	**62.37**	**75.54**
600	**86.65**	**77.79**	61.43	75.29

Table 5. Number of re-training cycles.

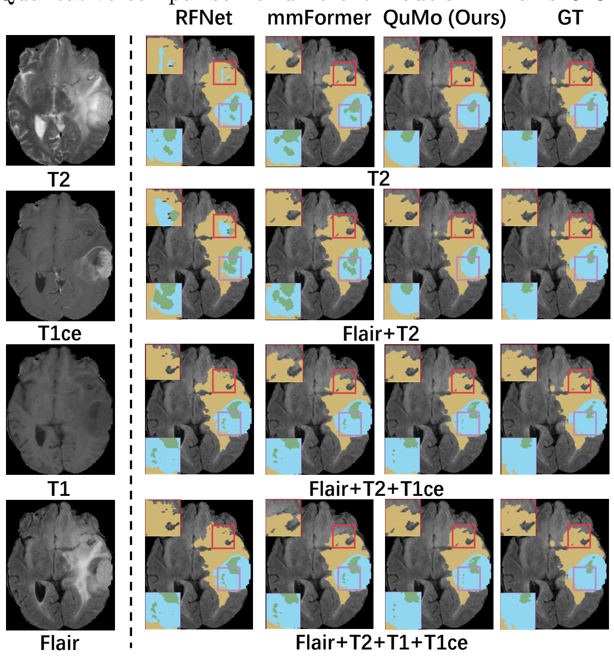
Cycle	Avg. Dice(%)	
	w/o R.T.	Ours
0	72.73	-
1	73.06	73.82
2	72.42	74.04
3	72.57	74.71
4	73.44	**75.54**
5	72.13	75.24

in our network. Then we add the MGT modules gradually on the Bottleneck Layer and the Skip Connection. We compare the performance of these networks on the Dice score, averaging over the 15 possible situations of input modalities. As shown in Table 3, we evaluate the influence of MGT in the bottleneck layer (Bottle.), skip-connection layers (Skip.), deep supervision (D.S.) and different initialization strategies (Random Initialization and Modal Code Initialization). Specifically, employing a randomly initialized MGT in the bottleneck layer without deep supervision increases the average Dice scores of three tumor regions by 3.92%, compared with " Baseline ", which demonstrates the superiority of the introduced queries. Moreover, our method of applying multi-scale MGTs with deep supervision increases the results over the " Baseline " by 4.45%.

As shown in Figure 5, we also visualized the attention maps corresponding to different queries of our proposed MGTs. The notation Q_n positioned on the left denotes the query's numerical index. Brighter areas represent greater activation values. It is evident from the figure that the sensitivity of the same query varies for different modal combinations, which way would make the decoder aware of the modal combinations.

Moreover, we analyze the impacts of the different numbers of queries. As shown in Table 4, performance increases with the number of queries until the number is around 500, since more queries contain more informative knowledge to perceive different modality-missing states.

Fig. 4. Qualitative comparison of different models in BraTs2018 dataset.

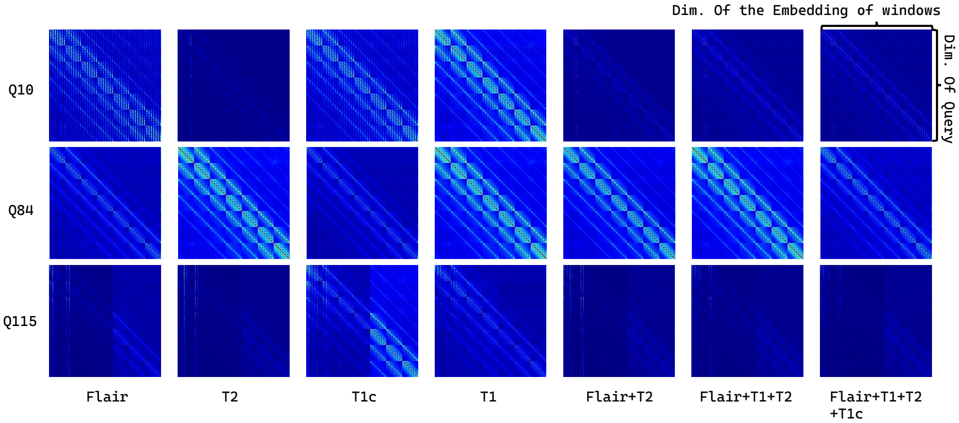


Furthermore, we investigate the impact of varying the number of re-training cycles on model performance. Specifically, a comparison was made between a training approach that did not incorporate the re-training strategy but had an equal total epoch count, and a training approach that incorporated the re-training strategy within each cycle, where the model weights for each cycle were initialized using the parameters generated by the preceding re-training cycle. The experimental outcomes are reported in Table 5, which exhibit a noteworthy enhancement in model performance following several re-training cycles. It was observed that a state of equilibrium was attained after four cycles, and therefore, four cycles were selected for the subsequent experiments.

Table 6. Comparison of computational complexity and model size.

Models	FLOPs(G)	Params(MB)
RFNet	830	**8.98**
mmFormer	<u>234</u>	35.34
MFI	2045	30.91
QuMo(Ours)	**233**	<u>24.65</u>

Fig. 5. Visualization of attention maps corresponding to different queries of our proposed MGTs. The notation Q_n positioned on the left denotes the query's numerical index. Brighter areas represent greater activation values. The same query is sensitive to different modal combinations and can be activated by different combinations in multiple degrees, in which way would make the decoder aware of the modal combinations. Furthermore, diagonal activation values, which are higher, indicating that these queries are proficient in acquiring location information, signifying that each part has the maximum response at its corresponding location.



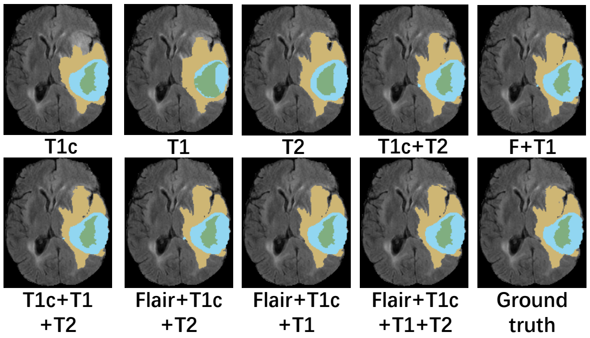
4 Conclusion

In this work, we design a novel incomplete multi-modal brain tumor segmentation method with a unified encoder-decoder architecture, which can provide the direct interaction of different modalities within the network and adopt a re-training mechanism to void convergence to sub-optimal local minima. Specifically, we apply the learnable modality combination embeddings (query) to guide the model to perceive different modality-missing states. Our model outperforms the state-of-the-art approach on the BraTS2018 and BraTS2020 datasets. However, despite the impressive performance of the QuMo, some additional works remain verified. We are particularly interested in QuMo's performance in other multi-modal tasks due to its availability for multi-modal perceptible interactions.

References

1. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: Proceedings of the European Conference on Computer Vision Workshops(ECCVW) (2022) 3
2. Carion, N., Massa, F., et al.: End-to-end object detection with transformers. In: Eur. Conf. Comput. Vis. pp. 213–229 (2020) 4
3. Chen, C., Dou, Q., Jin, Y., Chen, H., Qin, J., Heng, P.A.: Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd Inter-

Fig. 6. Visualization of the predicted segmentation results for various modality combinations.



national Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. pp. 447–456. Springer (2019) 2, 5, 6, 7, 8

4. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945) 6
5. Ding, Y., Yu, X., Yang, Y.: Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3975–3984 (2021) 2, 6, 7, 8
6. Dorent, R., Joutard, S., Modat, M., Ourselin, S., Vercauteren, T.: Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 74–82. Springer (2019) 6, 7, 8
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2020). https://doi.org/10.48550/ARXIV.2010.11929, https://arxiv.org/abs/2010.11929 2, 3
8. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021) 1
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proc. Int. Conf. Learn. Represent. (2015) 6
10. Liu, H., Fan, Y., Li, H., Wang, J., Hu, D., Cui, C., Lee, H.H., Zhang, H., Oguz, I.: Moddrop++: A dynamic filter network with intra-subject co-training for multiple sclerosis lesion segmentation with missing modalities (2022). https://doi.org/10.48550/ARXIV.2203.04959, https://arxiv.org/abs/2203.04959 2
11. Liu, Y., Fan, L., et al.: Incomplete multi-modal representation learning for alzheimer's disease diagnosis. Medical Image Analysis **69**, 101953 (2021) 2
12. Liu, Z., Lin, Y., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE Int. Conf. Comput. Vis. pp. 10012–10022 (2021) 3
13. Peiris, H., Hayat, M., Chen, Z., Egan, G., Harandi, M.: A robust volumetric transformer for accurate 3d tumor segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference,

Singapore, September 18–22, 2022, Proceedings, Part V. pp. 162–172. Springer (2022) 1, 3

14. Qiu, Y., Chen, D., Yao, H., Xu, Y., Wang, Z.: 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 (2023) 2
15. Qiu, Y., Zhao, Z., Yao, H., Chen, D., Wang, Z.: Modal-aware visual prompting for incomplete multi-modal brain tumor segmentation. In: Proceedings of the 31th ACM International Conference on Multimedia (2023) 2
16. Qu, M., Wu, Y., Liu, W., Gong, Q., Liang, X., Russakovsky, O., Zhao, Y., Wei, Y.: Siri: A simple selective retraining mechanism for transformer-based visual grounding. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV. pp. 546–562. Springer (2022) 2, 5
17. Tang, Y., Yang, D., Li, W., Roth, H., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis (2021). https://doi.org/10.48550/ARXIV.2111.14791, https://arxiv.org/abs/2111.14791 1
18. Valanarasu, J.M.J., Yasarla, R., et al.: Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In: IEEE Conf. Comput. Vis. Pattern Recognit. pp. 2353–2363 (2022) 2, 4
19. Vaswani, A., Shazeer, N., et al.: Attention is all you need. Proc. Adv. Neural Inf. Process. Syst. **30** (2017) 3, 4
20. Wang, L., Lee, C.Y., et al.: Training deeper convolutional networks with deep supervision. arXiv preprint arXiv:1505.02496 (2015) 5
21. Wang, S., et al.: Lt-net: Label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog. pp. 9162–9171 (2020) 2
22. Zhang, Y., He, N., et al.: mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. pp. 107–117 (2022) 2, 6, 7, 8
23. Zhang, Y., Yang, J., Tian, J., Shi, Z., Zhong, C., Zhang, Y., He, Z.: Modality-aware mutual learning for multi-modal medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 589–599. Springer (2021) 1
24. Zhao, Z., Yang, H., et al.: Modality-adaptive feature interaction for brain tumor segmentation with missing modalities. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. pp. 183–192 (2022) 2
25. Zhou, C., Ding, C., Lu, Z., Wang, X., Tao, D.: One-pass multi-task convolutional neural networks for efficient brain tumor segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11. pp. 637–645. Springer (2018) 1