NestedFormer: Nested Modality-Aware Transformer for Brain Tumor Segmentation

Zhaohu Xing[1], Lequan Yu[2], Liang Wan[1(✉)], Tong Han[3], and Lei Zhu[4,5]

[1] Medical College of Tianjin University, Tianjin, China
xingzhaohu@tju.edu.cn, lwan@tju.edu.cn
[2] The University of Hong Kong, Hong Kong, China
[3] Brain Medical Center of Tianjin University, Huanhu Hospital, Tianjin, China
[4] The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[5] The Hong Kong University of Science and Technology, Hong Kong, China

Abstract. Multi-modal MR imaging is routinely used in clinical practice to diagnose and investigate brain tumors by providing rich complementary information. Previous multi-modal MRI segmentation methods usually perform modal fusion by concatenating multi-modal MRIs at an early/middle stage of the network, which hardly explores non-linear dependencies between modalities. In this work, we propose a novel Nested Modality-Aware Transformer (NestedFormer) to explicitly explore the intra-modality and inter-modality relationships of multi-modal MRIs for brain tumor segmentation. Built on the transformer-based multi-encoder and single-decoder structure, we perform nested multi-modal fusion for high-level representations of different modalities and apply modality-sensitive gating (MSG) at lower scales for more effective skip connections. Specifically, the multi-modal fusion is conducted in our proposed Nested Modality-aware Feature Aggregation (NMaFA) module, which enhances long-term dependencies within individual modalities via a tri-orientated spatial-attention transformer, and further complements key contextual information among modalities via a cross-modality attention transformer. Extensive experiments on BraTS2020 benchmark and a private meningiomas segmentation (MeniSeg) dataset show that the NestedFormer clearly outperforms the state-of-the-arts. The code is available at https://github.com/920232796/NestedFormer.

Keywords: Multi-modal MRI · Brain Tumor Segmentation · Nested Modality-Aware Feature Aggregation · Modality-Sensitive Gating

1 Introduction

Brain tumor is one of the most common cancers in the world [3], in which gliomas are the most common malignant brain tumors with different levels of aggressiveness and meningiomas are the most prevalent primary intracranial tumors in adults [21]. Multi-modal magnetic resonance imaging (MRI) is routinely used in the clinic by providing rich complementary information for analyzing brain
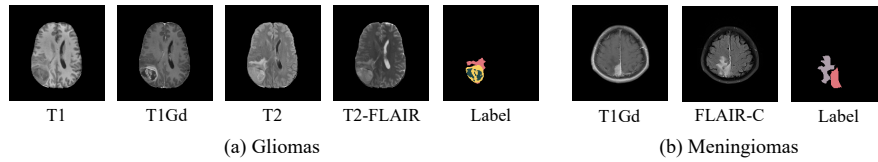
Fig. 1: Multi-modal MRIs for (a) Gliomas; and (b) Meningiomas.

tumors. Specifically, for gliomas, the commonly used MRI sequences are T1-weighted (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2) and T2 Fluid Attenuation Inversion Recovery (T2-FLAIR) images; see Fig. 1(a), each with varying roles in distinguishing tumor, peritumoral edema and tumor core [1, 2, 18]. For meningiomas, they have different characteristic appearances on T1Gd [15] and contrast-enhanced T2-FLAIR (shorted for FLAIR-C) MRI images; see Fig. 1(b). Thus, automatic segmentation of brain tumor structures from multi-modal MRIs is important for clinical diagnosis and treatment planning.

In recent years, convolutional neural networks (CNNs) have achieved promising successes in brain tumor segmentation. The main stream models are built upon the encoder-decoder architecture [22] with skip connections, including S3D-UNet [5], SegResNet [20], HPU-Net [13], etc. Recent works [9,25,28] also explore transformer [24] to model long-range dependencies within images. For instance, TransBTS [25] utilizes 3D-CNN to extract local spatial features, and applies transformer to model global dependencies on high-level features. UNETR [9] uses the ViT transformer as the encoder to learn contextual information, which is merged with the CNN-based decoder via skip connections at multiple resolutions. However, the transformer in these methods is used to enhance the encode path without specific design for multi-modal fusion.

To utilize the multi-modal information, most of existing methods adopt an early-fusion strategy, in which multi-modal images are concatenated as the network input. However, this strategy can hardly explore non-linear relationships between different modalities. To alleviate this problem, recent works follow a layer-fusion strategy [7, 27, 29], where the modality-specific features extracted by different encoders are fused in the middle layers of the network and share the same decoder. In HyeprDense-Net [7], each modality has a separated stream and dense connections are introduced between layers within the same stream and also across different streams. MAML [27] embeds multi-modal images by different modality-specific FCNs and then applies a modality-aware module to regress attention maps in order to fuse the modality-specific features. Nevertheless, these multi-modal fusion methods do not build the long-range spatial dependencies within and cross modalities, so that they cannot fully utilize the complementary information of different modalities.

In this paper, we propose a novel nested modality-aware transformer, called NestedFormer, for effective and robust multi-modal brain tumor segmentation. We first design an effective Global Poolformer to extract discriminative volumetric spatial features, with more emphasis on global dependencies, from different
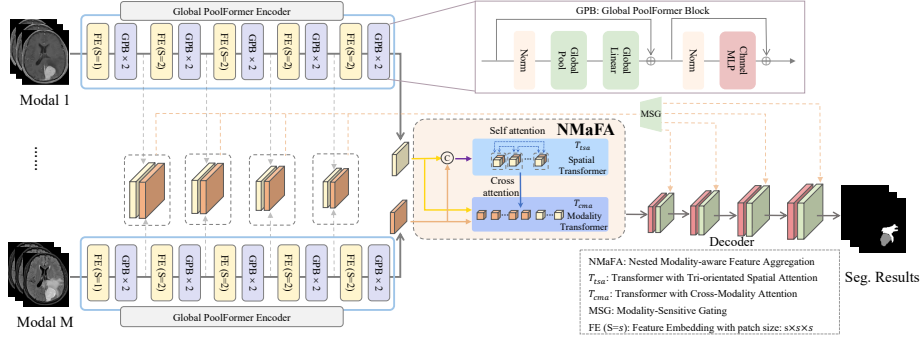
Fig. 2: An overview of the proposed NestedFormer. We design a Nested Modality-aware Feature Aggregation (NMaFA) module to model both the intra- and inter-modality features for multi-modal fusion.

MRI modalities. To better extract the complementary features and enable any number of modalities for fusion, we propose a novel Nested Modality-aware Feature Aggregation (NMaFA) module. It explicitly considers both single-modality spatial coherence and cross-modality coherence, and leverages nested transformers to establish the intra- and inter-modality long-range dependencies, resulting in more effective feature representation. Moreover, we design a computationally efficient Tri-orientated Spatial Attention (TSA) paradigm to accelerate the 3D-spatial-coherence calculation. To improve feature reuse effect in the decoding, a novel modality-sensitive gating (MSG) module is developed to dynamically filter modality-aware low-resolution features for effective skip connections. Extensive experiments on BraTS2020 benchmark and a privately collected meningiomas segmentation dataset (MeniSeg) show that our model clearly ourperforms the state-of-the-art methods.

2   Method

Fig. 2 illustrates the overview of the proposed NestedFormer, which consists of three components: 1) multiple encoders to obtain multi-scale representations of different modalities, 2) a NMaFA fusion module to explore correlated features within and between multi-modal high-level embeddings, and 3) a gating strategy to selectively transfer modality-sensitive low-resolution features to the decoder.

2.1   Global Poolformer Encoder

Recent works show that transformer is more conducive to modeling global information than CNNs. To better extract local context information for each modality, we extend the Poolformer [26] as the modality-specific encoder. As discussed in [26], replacing the computation-intensive attention module in Transformer with average pooling can achieve superior performance than recent transformer and MLP-like models. Therefore, to enhance global information, we design Global Poolformer Block (GPB), which leverages global pooling instead of

average pooling in Poolformer, followed by a fully connected layer. As shown in Fig. 2, given the input feature embedding X, a GPB block consists of a learnable global pooling (GP) and a MLP sub-block. The output Z is computed as,

$$Y = GP(LN(X))W_g + X,$$
$$Z = MLP(LN(Y)) + Y, \tag{1}$$

where $LN(*)$ denotes the layer normalization and W_g is the learnable parameter in the FC layer. Our Global Poolformer encoder contains five groups of one feature embedding (FE) layer and two GPB blocks. Each FE layer is a 3D-convolution, while the first FE layer has a convolution patch size of $1 \times 1 \times 1$ and the rest layers have a patch size of $2 \times 2 \times 2$ and a stride of 2. The encoders gradually encode each modality image into high-level feature $F_{L,i} \in \mathbb{R}^{d \times w \times h \times C}, i \in [1, M]$, where $(d, w, h) = (\frac{D}{16}, \frac{W}{16}, \frac{H}{16})$ are 1/16 of input spatial resolutions H, W and depth dimension D; M is the number of modal images, the channel dimension C and the layer number L are set as $C = 128$, $L = 5$.

2.2   Nested Modality-Aware Feature Aggregation

Given high-level features $F_{L,i}, i \in [1, M]$, NMaFA leverages a spatial-attention based transformer T_{tsa} and a cross-modality attention based transformer T_{cma} in a nested manner; see Fig. 3. First, transformer T_{tsa} utilizes the self-attention scheme to compute the long-range correlation between different patches in the space within each modality. Specifically, $F_{L,i}$ is *concatenated in the channel dimension* to obtain high-level embedding $F_s \in \mathbb{R}^{d \times w \times h \times MC}$. In this work, each location of F_s is considered as one "patch". Then a patch embedding layer maps F_s to a token sequence $\hat{F}_s \in \mathbb{R}^{dwh \times C}$. T_{tsa} takes $\hat{F}_s$ and the position encoding [24] as the input, and outputs spatially-enhanced feature $\tilde{F}_s \in \mathbb{R}^{dwh \times C}$.

Second, transformer T_{cma} utilizes the cross-attention scheme to further compute the global relation among different modalities to achieve inter-modality fusion. To this end, $F_{L,i}$ is *concatenated in the spatial dimension* to obtain the flatten sequence $\hat{F}_c \in \mathbb{R}^{MP \times C}$. Here, $P(P = 32)$ denotes the number of dominant tokens learnt via the Token Learner strategy [23], which helps to reduce the computational scope especially when the number of tokens increases greatly along with more modalities. After that, both $\tilde{F}_s$ and $\hat{F}_c$ are fed into T_{cma} to get the modality-enhanced feature embedding $\tilde{F}$.

Also note that our two modules are different from traditional channel-spatial attention networks, which reweigh feature maps channel-wise and spatial-wise. Our NMaFA relies on transformer mechanism and the two transformers are fused in a nested form, rather than serial [12] or parallel [19] fusion.

Transformer with Tri-orientated Spatial Attention. To improve the computational efficiency of spatial attention for volumetric embeddings, inspired by [10,16], we leverage axial-wise attention MHA_z, plane-wise attention MHA_{xy}, and window-wise attention MHA_w; see Fig. 3(b). Concretely, MHA_z models the long-range relationship among feature tokens along the vertical direction; MHA_{xy} models the long-range relationship within each slice; MHA_w uses sliding windows to model the relationship across local 3D-windows. We employ
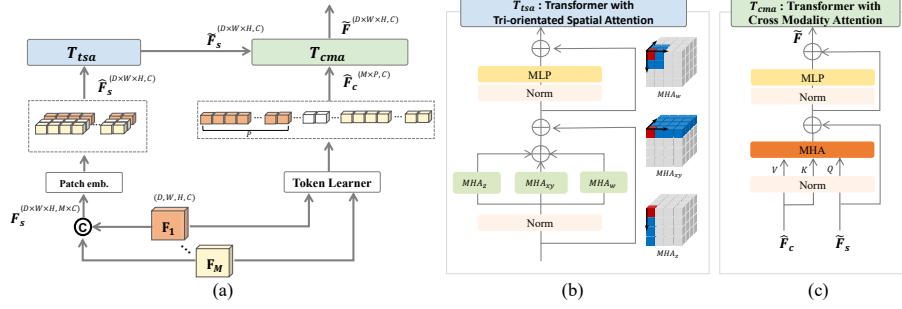
Fig. 3: NMaFA: Nested Modality-aware Feature Aggregation. (a) The overall architecture. (b) The transformer with tri-orientated spatial attention T_{tsa}. (c) The transformer with cross-modality attention T_{cma}.

axial and planar learnable absolute position encodings [24] for MHA_z and MHA_{xy}, respectively, and use relative position encoding for window-wise attention MHA_w [16]. The resultant attention is computed as follows,

$$MHA_{tsa}(z) = MHA_z(z) + MHA_{xy}(z) + MHA_w(z), \tag{2}$$

where $z \in \mathbb{R}^{N \times C}$ denotes the embedding tokens with sequence length N and embedding dimension C after layer normalization, $N = d \times w \times h$. By this way, the model not only enhances feature extraction of local important regions, but also calculates global feature dependencies with less computation.

Transformer with Cross-Modality Attention. By concatenating features *in the channel dimension*, T_{tsa} mainly enhances the dependencies within each modality and yields $\tilde{F}_s$, although the inter-modality integration also takes place via patch embedding. To explicitly explore relationship among modalities, we concatenate the feature tokens of different modalities *along the spatial dimension*, yielding $\hat{F}_c \in \mathbb{R}^{MP \times C}$; and then use a cross-attention transformer T_{cma} to enhance the modality dependency information into $\tilde{F}_s$; see Fig. 3(c). The input triplet of (Query, Key, Value) to the cross-attention is computed as

$$Q = \tilde{F}_s W_q, \ K = \hat{F}_c W_k, \ V = \hat{F}_c W_v, \tag{3}$$

where $W_q, W_k, W_v \in \mathbb{R}^{C \times d}$ are the weight matrices, $d = 128$ is the dimension of Q, K, V. The cross attention CA is then formulated as

$$CA(\tilde{F}_s, \hat{F}_c) = \hat{F}_s + SoftMax(\frac{QK^T}{\sqrt{d}})V. \tag{4}$$

The resultant token sequence $\tilde{F}$ from T_{cma} fuses and enhances the input features with increasing receptive fields and the cross-modal global relevance.

2.3   Modality-Sensitive Gating

In feature decoding, we first fold the tokens $\tilde{F}$ back to a high-level 4D feature map $R_L \in \mathbb{R}^{d \times w \times h \times C}$. R_L is progressively processed in a regular bottom-up style

with a 3D convolution and $2\times$ upsampling operation to recover a full resolution feature map $R_1 \in \mathbb{R}^{D \times W \times H \times N_c}$ for segmentation, where N_c is the number of segments. Note that the encoder features are multi-modal. Hence, we design a modality-sensitive gating strategy in skip connection, to filter the encoder features $\{F_{l,i}, l \in [1,4], i \in [1, M]\}$ according to the modality importance. To be specific, for the l-th layer, an modality importance map $I_l \in \mathbb{R}^{\frac{D}{2^{l-1}} \times \frac{W}{2^{l-1}} \times \frac{H}{2^{l-1}} \times M}$ is learnt from $\tilde{F}$ that is the output of NMaFA, as follows,

$$I_l = \sigma(U_{L-l}^{2\times}(FC(\tilde{F}))), \tag{5}$$

where $FC(*)$ is a $1 \times 1 \times 1$ full connection layer, $U_l^{2\times}$ denotes upsampling l times, and $\sigma(*)$ is the sigmoid function. Denote $\odot$ as element-wise multiplication. Then the filtered encoder feature $\mathcal{F}_l$ is formulated as

$$\mathcal{F}_l = \sum_i (I_{l,i} \odot F_{l,i}). \tag{6}$$

3   Experiment

3.1   Implementation Details

Our NestedFormer was implemented in PyTorch1.7.0 on a NVIDIA GTX 3090 GPU. The parameters were initialized via Xavier [8]. The loss function was a combination of soft dice loss and cross-entropy loss and we adopted the AdamW optimizer [17] with a weight decay of 10^{-5}. The learning rate was empirically set as 10^{-4}.We adopted two T_{tsa} sequentially and just one T_{cma}. In MHA_w, the window-size was set as (2, 2, 2) for BraTS2020 and (2, 4, 4) for MeniSeg.

3.2   Datasets and Evaluation Metrics

For evaluation, we use a public brain tumor segmentation dataset **BraTS2020** [18] and a private 3D meningioma segmentation dataset (**MeniSeg**) collected from Brain Medical Center of Tianjin University, Tianjin Huanhu hospital. Dice score and 95% Hausdorff Distance (HD95) are adopted for quantitative comparison.

BraTS2020 Dataset.   The BraTS2020 training dataset contains 369 aligned four-modal MRI data (i.e., T1, T1Gd, T2, T2-FLAIR), with expert segmentation masks (i.e., GD-enhancing tumor, peritumoral edema, and tumor core). Each modality has a $155 \times 240 \times 240$ volume and is already resampled and co-registered. The segmentation task aims to segment the whole tumor (WT), enhancing tumor (ET), and tumor core (TC) regions. Following the recent work [14], we randomly divide the dataset into training (315), validation (17) and test (37).

Meningioma Dataset.   The MeniSeg dataset contains 110 annotated two-modal MRIs (i.e., T1Gd and FLAIR-C) from the meningiomas patients, who had undergone tumor resection between March 2016 and March 2021. MRI scans were performed with four 3.0T MRI scanners (Skyra, Trio, Avanto, Prisma from Siemens). Two radiologists annotated meningioma tumor and edema masks on T1Gd and FLAIR-C MRIs, and the third high-experienced radiologist made examination. Each modality data has a volume of $32 \times 256 \times 256$, and is aligned into the same space and sampled to volume sizes of [32, 192, 192] for training. Two-fold cross-validation is conducted for all the compared methods.

Table 1: Quantitative comparison on BraTS 2020 dataset.

Methods	Param (M)	FLOPs (G)	WT		TC		ET		Ave	
			Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓
3D-UNet [6]	5.75	1449.59	0.882	5.113	0.830	6.604	0.782	6.715	0.831	6.144
SegResNet [20]	18.79	185.23	0.903	4.578	0.845	5.667	0.796	7.064	0.848	5.763
MAML [27]	5.76	577.65	0.914	4.804	0.854	5.594	0.796	**5.221**	0.855	5.206
nnUNet [11]	5.75	1449.59	0.907	6.94	0.848	5.069	**0.814**	5.851	0.856	5.953
SwinUNet(2D) [4]	27.17	357.49	0.872	6.752	0.809	8.071	0.744	10.644	0.808	8.489
TransBTS [25]	32.99	333	0.910	**4.141**	0.855	5.894	0.791	5.463	0.852	5.166
UNETR [9]	92.58	41.19	0.899	4.314	0.842	5.843	0.788	5.598	0.843	5.251
NestedFormer	10.48	71.77	**0.920**	4.567	**0.864**	**5.316**	0.800	5.269	**0.861**	**5.051**

3.3 Comparison with SOTA Methods

We compare our network against seven SOTA segmentation methods, including three CNN-based methods (3D-UNet [6], SegResNet [20], MAML [27], nnUNet [11]), and three transformer-based methods (SwinUNet(2D) [4], TransBTS [25], and UNETR [9]). For a fair comparison, we utilize the public implementations of compared methods to re-train their networks for generating their best segmentation results. Considering the computation power, all the methods are trained for at most 300 epochs on BraTS2020 and 200 epochs on MeniSeg.

BraTS2020. Table 1 reports the Dice and HD95 scores on three regions (WT, TC, and ET) as well as the averaged scores of all the methods on BraTS2020. Apparently, our NestedFormer achieves the largest Dice score on WT, the largest Dice score on TC, the smallest HD95 scores on TC, and our method also ranks second in Dice score on ET, and second in HD95 score on WT and ET. More importantly, our method has the best quantitative performance with averaging Dice and HD95 scores to be 0.861 and 5.051. It is noted that HD95 is for the distance difference between two sets of points, which is more sensitive than Dice [25]. Hence, Dice is often used as the main metric and HD95 as the reference. We also experimented with two-fold cross-validation for UNETR, TransBTS and our method, while our method outperforms the two methods in WT and TC, and is quite close to the best result in ET. As for model complexity, our model has 10.48M parameters and 71.77G FLOPs which is a moderate size model.

MeniSeg. In Table 2, we list Dice and HD95 scores of our network and compared methods on tumor and edema regions on the MeniSeg dataset, as well as the average metrics. Among all the compared methods, MAML has the largest Dice score of 0.819 at the tumor segmentation, while UNETR has the largest Dice score of 0.693 at the edema segmentation, and average Dice score of 0.755. In comparison, our method has a 1.5% Dice improvement in meningioma tumor, 0.2% Dice improvement in edema, and 1.0% average Dice improvement. Regarding HD95, our method achieves the 4th smallest score of 2.647 on the tumor segmentation, and the smallest score of 6.173 on the edema segmentation.

Visual Comparisons on BraTS2020 and MeniSeg. Fig. 4 visually compares the segmentation results predicted by our network and SOTA methods on BraTS2020 and MeniSeg. From these visualization results, we can find that our method can more accurately segment brain tumor and peritumoral edema

Table 2: Quantitative comparison on MeniSeg dataset.

Methods	Tumor		Edema		Ave	
	Dice	HD95	Dice	HD95	Dice	HD95
3D-UNet [6]	0.799	5.099	0.676	9.655	0.737	7.377
SegResNet [20]	0.813	2.970	0.665	10.438	0.739	6.704
MAML [27]	0.819	2.112	0.682	9.158	0.750	5.635
SwinUNet(2D) [4]	0.807	1.817	0.679	7.986	0.743	4.901
TransBTS [25]	0.809	**1.742**	0.679	6.388	0.744	**4.065**
UNETR [9]	0.818	3.279	0.693	7.837	0.755	5.813
NestedFormer	**0.834**	2.647	**0.695**	**6.173**	**0.765**	4.410

Table 3: Ablation study for different modules on MeniSeg.

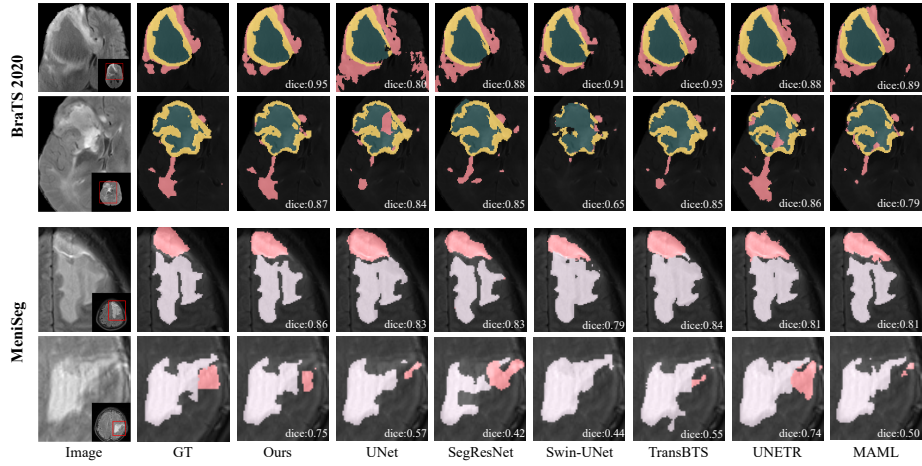| Methods	Encoder			Fusion			Dice		
	CNN	PB	GPB	T_{tsa}	T_cma	MSG	Tmuor	Edema	Ave
baseline 1	✓						0.805	0.675	0.74
baseline 2			✓				0.810	0.679	0.75
baseline 3			✓	✓			0.816	0.688	0.752
baseline 4			✓	✓	✓		0.825	**0.699**	0.762
baseline 5		✓		✓	✓	✓	0.823	0.697	0.76
NestedFormer			✓	✓	✓	✓	**0.834**	0.695	**0.765**

Fig. 4: The visual comparison results on BraTS2020 and MeniSeg dataset.

regions than all the compared methods. The reason behind is that our method is able to better fuse multi-modal MRIs by explicitly exploring the intra-modality and the inter-modality relationships among multiple modalities.

3.4   Ablation study

We conduct ablation studies on the MeniSeg dataset to evaluate the contributions of main modules in our method; see Table 3. We not only compared the effects of three different encoder backbones based on CNN, PB, and GP, but also verified the effect of our proposed fusion modules. Among them, $baseline1$ uses multiple U-Net encoders to extract features of different modal images, and performs feature fusion by concatenation. $baseline2$-$baseline4$ uses multiple GPB encoders to extract features, and conducts skip connection via simple convolution, w/o T_{tsa} and T_{cma} (see Fig. 3), respectively. $baseline5$ replaces GPB block with the original PoolFormer block (PB) in the encoder, using the proposed NMaFA module (including T_{tsa} and T_{cma}) as well as MSG. It can be observed clearly that compared with $baseline2$, using the NMaFA module enhances the

extraction of long-distance dependency information and effectively improves the segmentation results, while GPB outperforms PB by considering global information. Moreover, the MSG module is added to increase the feature reuse capability of skip connections, which further improves the segmentation effect, achieving the best average segmentation Dice (0.765) on the MeniSeg dataset.

4   Conclusion

We propose a novel multi-modal segmentation framework, dubbed as Nested-Former. This architecture extracts the features of M modalities by using multiple Global Poolformer Encoders. Then, the high-level features are effectively fused by the NMaFA module, and the low-level features are selected by the modality-sensitive gate (MSG) module. Through these proposed modules, the network effectively extracts and hierarchically fuses features from different modalities. The effectiveness of our proposed NestedFormer is validated on BraTS2020 and MeniSeg datasets. Our framework are modality-agnostic and can be extended to other multimodal medical data. In the future work, we will explore more efficient feature fusion on low-levels to further improve the segmentation performance.

References

1. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific Data **4**(1), 1–13 (2017)
2. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., 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)
3. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R., Torre, L., Jemal, A.: Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians **68**(6), 394–424 (2018)
4. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021)
5. Chen, W., Liu, B., Peng, S., Sun, J., Qiao, X.: S3d-unet: separable 3d u-net for brain tumor segmentation. In: International MICCAI Brainlesion Workshop. pp. 358–368. Springer (2018)
6. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 424–432. Springer (2016)

7. Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C., Ayed, I.B.: Hyperdense-net: A hyper-densely connected cnn for multi-modal image segmentation. IEEE Transactions on Medical Imaging **38**(5), 1116–1126 (2019)
8. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256. JMLR Workshop and Conference Proceedings (2010)
9. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584 (2022)
10. Ho, J., Kalchbrenner, N., Weissenborn, D., Salimans, T.: Axial attention in multi-dimensional transformers. arXiv preprint arXiv:1912.12180 (2019)
11. 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)
12. Khanh, T.L.B., Dao, D.P., Ho, N.H., Yang, H.J., Baek, E.T., Lee, G., Kim, S.H., Yoo, S.B.: Enhancing u-net with spatial-channel attention gate for abnormal tissue segmentation in medical imaging. Applied Sciences **10**(17), 5729 (2020)
13. Kong, X., Sun, G., Wu, Q., Liu, J., Lin, F.: Hybrid pyramid u-net model for brain tumor segmentation. In: International Conference on Intelligent Information Processing. pp. 346–355. Springer (2018)
14. Larrazabal, A.J., Martínez, C., Dolz, J., Ferrante, E.: Orthogonal ensemble networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 594–603. Springer (2021)
15. Li, X., Lu, Y., Xiong, J., Wang, D., She, D., Kuai, X., Geng, D., Yin, B.: Presurgical differentiation between malignant haemangiopericytoma and angiomatous meningioma by a radiomics approach based on texture analysis. Journal of Neuroradiology **46**(5), 281–287 (2019)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
18. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2014)
19. Mou, L., Zhao, Y., Chen, L., Cheng, J., Gu, Z., Hao, H., Qi, H., Zheng, Y., Frangi, A., Liu, J.: Cs-net: channel and spatial attention network for curvilinear structure segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 721–730. Springer (2019)
20. Myronenko, A.: 3d mri brain tumor segmentation using autoencoder regularization. In: International MICCAI Brainlesion Workshop. pp. 311–320. Springer (2018)
21. Ostrom, Q.T., Patil, N., Cioffi, G., Waite, K., Kruchko, C., Barnholtz-Sloan, J.S.: Cbtrus statistical report: primary brain and central nervous system tumors diagnosed in the united states in 2013–2017. Neuro Oncology **22**(iv), 1–96 (2020)
22. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)

23. Ryoo, M.S., Piergiovanni, A., Arnab, A., Dehghani, M., Angelova, A.: Token-learner: What can 8 learned tokens do for images and videos? arXiv preprint arXiv:2106.11297 (2021)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
25. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: Transbts: Multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 109–119. Springer (2021)
26. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: Metaformer is actually what you need for vision. arXiv preprint arXiv:2111.11418 (2021)
27. 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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 589–599. Springer (2021)
28. Zhang, Y., Pei, Y., Zha, H.: Learning dual transformer network for diffeomorphic registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 129–138. Springer (2021)
29. Zhou, T., Canu, S., Vera, P., Ruan, S.: 3d medical multi-modal segmentation network guided by multi-source correlation constraint. In: 25th International Conference on Pattern Recognition. pp. 10243–10250. IEEE (2020)