mmFormer: Multimodal Medical Transformer for Incomplete Multimodal Learning of Brain Tumor Segmentation

Yao Zhang[1,2], Nanjun He[3], Jiawei Yang[4], Yuexiang Li[3], Dong Wei[3],
Yawen Huang[3(✉)], Yang Zhang[5], Zhiqiang He[5(✉)], and Yefeng Zheng[3]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
zhangyao215@mails.ucas.ac.cn
[3] Jarvis Lab, Tencent, Shenzhen, China
yawenhuang@tencent.com
[4] Electrical and Computer Engineering, University of California, Los Angeles, USA
[5] Lenovo Research, Beijing, China

Abstract. Accurate brain tumor segmentation from Magnetic Resonance Imaging (MRI) is desirable to joint learning of multimodal images. However, in clinical practice, it is not always possible to acquire a complete set of MRIs, and the problem of missing modalities causes severe performance degradation in existing multimodal segmentation methods. In this work, we present *the first attempt* to exploit the Transformer for multimodal brain tumor segmentation that is robust to any combinatorial subset of available modalities. Concretely, we propose a novel multi**m**odal **M**edical Trans**former** (**mmFormer**) for *incomplete multimodal learning* with three main components: the hybrid modality-specific encoders that bridge a convolutional encoder and an intra-modal Transformer for both local and global context modeling within each modality; an inter-modal Transformer to build and align the long-range correlations across modalities for modality-invariant features with global semantics corresponding to tumor region; a decoder that performs a progressive up-sampling and fusion with the modality-invariant features to generate robust segmentation. Besides, auxiliary regularizers are introduced in both encoder and decoder to further enhance the model's robustness to incomplete modalities. We conduct extensive experiments on the public BraTS 2018 dataset for brain tumor segmentation. The results demonstrate that the proposed mmFormer outperforms the state-of-the-art methods for incomplete multimodal brain tumor segmentation on almost all subsets of incomplete modalities, especially by an average 19.07% improvement of Dice on tumor segmentation with only one available modality. The code is available at https://github.com/YaoZhang93/mmFormer.

Keywords: Incomplete multimodal learning · Brain tumor segmentation · Transformer

1   Introduction

Automated and accurate segmentation of brain tumors plays an essential role in clinical assessment and diagnosis. Magnetic Resonance Imaging (MRI) is a common neuroimaging technique for the quantitative evaluation of brain tumors in clinical practice, where multiple imaging modalities, i.e., T1-weighted (T1), contrast-enhanced T1-weighted (T1c), T2-weighted (T2), and Fluid Attenuated Inversion Recovery (FLAIR) images, are provided. Each imaging modality provides a distinctive contrast of the brain structure and pathology. The joint learning of multimodal images for brain tumor segmentation is essential and can significantly boost the segmentation performance. Plenty of methods have been widely explored to effectively fuse multimodal MRIs for brain tumor segmentation by, for example, concatenating multimodal images in channel dimension as the input or fusing features in the latent space [17,23]. However, in clinical practice, it is not always possible to acquire a complete set of MRIs due to data corruption, various scanning protocols, and unsuitable conditions of patients. In this situation, most existing multimodal methods may fail to deal with incomplete imaging modalities and face a severe degradation in segmentation performance. Consequently, a robust multimodal method is highly desired for a flexible and practical clinical application with one or more missing modalities.

Incomplete multimodal learning, also known as hetero-modal learning [8], aims at designing methods that are robust with any subset of available modalities at inference. A straightforward strategy for incomplete multimodal learning of brain tumor segmentation is synthesizing the missing modalities by generative models [18]. Another stream of methods explores knowledge distillation from complete modalities to incomplete ones [2,10,21]. Although promising results are obtained, such methods have to train and deploy a specific model for each subset of missing modalities, which is complicated and burdensome in clinical application. Zhang et al. [22] proposed an ensemble learning of single-modal models with adaptive fusion to achieve multimodal segmentation. However, it only works when one or all modalities are available. Meanwhile, all these methods require complete modalities during the training process.

Recent methods focused on learning a unified model, instead of a bunch of distilled networks, for incomplete multimodal segmentation [8,16]. For example, HeMIS [8] learns an embedding of multimodal information by computing mean and variance across features from any number of available modalities. U-HVED [4] further introduces multimodal variational auto-encoder to benefit incomplete multimodal segmentation with generation of missing modalities. More recent methods also proposed to exploit feature disentanglement [1] and attention mechanism [3] for robust multimodal brain tumor segmentation. Fully Convolutional Network (FCN) [11,15] has achieved great success in medical image segmentation and is widely used for feature extraction in the methods mentioned above. Despite its excellent performance, the inductive bias of convolution, i.e., the locality, makes FCN difficult to build long-range dependencies explicitly. In incomplete multimodal learning of brain tumor segmentation, the features extracted with limited receptive fields tend to be biased when dealing with varying modalities.
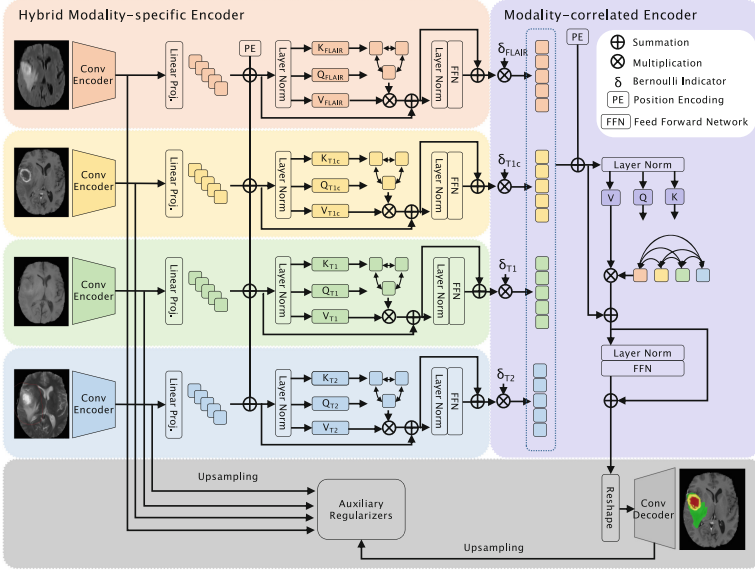
Fig. 1. Overview of the proposed mmFormer, which is composed of four hybrid modality-specific encoders, a modality-correlated encoder, and a convolutional decoder. Meanwhile, auxiliary regularizers are introduced in both encoder and decoder. The skip connections between the convolutional encoder and decoder are hidden for clear display.

In contrast, a modality-invariant embedding with global semantic information of tumor region across different modalities may contribute to more robust segmentation, especially when one or more modalities are missing.

Transformer was originally proposed to model long-range dependencies for sequence-to-sequence tasks [19], and also shows state-of-the-art performance on various computer vision tasks [5]. Concurrent works [7,14,20] exploited Transformer for brain tumor segmentation from the view of backbone network. However, the dedicated Transformer for multimodal modeling of brain tumor segmentation has not been carefully tapped yet, letting alone the incomplete multimodal segmentation.

This paper aims to exploit Transformer to build a unified model for incomplete multimodal learning of brain tumor segmentation. We propose Multimodal **Medical Transformer** (**mmFormer**) that leverages hybrid modality-specific encoders and a modality-correlated encoder to build the long-range dependencies both within and across different modalities. With the modality-invariant representations extracted by explicitly building and aligning global correlations between different modalities, the proposed mmFormer demonstrates superior robustness to incomplete multimodal learning of brain tumor segmentation. Meanwhile, auxiliary regularizers are introduced into mmFormer to encourage both encoder and decoder to learn discriminative features even when a certain number of modalities are missing. We validate mmFormer on the task of

multimodal brain tumor segmentation with BraTS 2018 dataset [12]. The proposed method outperforms the state-of-the-art methods in the average Dice metric over all settings of missing modalities, especially by an average 19.07% improvement in Dice on enhancing tumor segmentation with only one available modality. To the best of our knowledge, *this is the first attempt to involve the Transformer for incomplete multimodal learning of brain tumor segmentation.*

2   Method

In this paper, we propose mmFormer for incomplete multimodal learning of brain tumor segmentation. We adopt an encoder-decoder architecture to construct our mmFormer, including a hybrid modality-specific encoder for each modality, a modality-correlated encoder, and a convolutional decoder. Besides, auxiliary regularizers are introduced in both encoder and decoder. An overview of mmFormer is illustrated in Fig. 1. We elaborate on the details of each component in the followings.

2.1   Hybrid Modality-Specific Encoder

The hybrid modality-specific encoder aims to extract both local and global context information within a specific modality by bridging a convolutional encoder and an intra-modal Transformer. We denote the complete set of modalities by $M = \{FLAIR, T1c, T1, T2\}$. Given an input of $\mathbf{X}_m \in \mathbb{R}^{1 \times D \times H \times W}$ with a size of $D \times H \times W$, $m \in M$, we first utilize the convolutional encoder to generate compact feature maps with the local context and then leverage the intra-modal Transformer to model the long-range dependency in a global space.

Convolutional Encoder. The convolutional encoder is constructed by stacking convolutional blocks, similar to the encoder part of U-Net [15]. The feature maps with the local context within each modality produced by the convolutional encoder $\mathcal{F}_m^{conv}$ can be formulated as

$$\mathbf{F}_m^{local} = \mathcal{F}_m^{conv}(\mathbf{X}_m; \theta_m^{conv}) \tag{1}$$

where $\mathbf{F}_m^{local} \in \mathbb{R}^{C \times \frac{D}{2^{l-1}} \times \frac{H}{2^{l-1}} \times \frac{W}{2^{l-1}}}$, C is the channel dimension, and l is the number of the stages in the encoder. Concretely, we build a five-stage encoder, and each stage consists of two convolutional blocks. Each block contains cascaded group normalization, ReLU, and convolutional layers with kernel size of 3, while the first convolutional block in the first stage only contains a convolutional layer. Between two consecutive blocks, a convolutional layer with stride of 2 is employed to downsample the feature maps. The number of filters at each level of the encoder is 16, 32, 64, 128, and 256, respectively.

Intra-modal Transformer. Limited by the intrinsic locality of the convolutional network, the convolutional encoder fails to effectively build the long-range dependency within each modality. Therefore, we exploit the Intra-modal Transformer for explicitly long-range contextual modeling. The Intra-modal Transformer contains a tokenizer, a Multi-head Self Attention (MSA), and a Feed-Forward Network (FFN).

As Transformer processes the embeddings in a sequence-to-sequence manner, the local feature maps $\mathbf{F}_m^{local}$ produced by the convolutional encoder is first flattened into a 1D sequence and transformed into token space by a linear projection. However, the flattening operation inevitably collapses the spatial information, which is critical to image segmentation. To address this issue, we introduce a learnable position embedding $\mathbf{P_m}$ to supplement the flattened features via element-wise summation, which is formulated as

$$\mathbf{F}_m^{token} = \mathbf{F}_m^{local}\mathbf{W}_m + \mathbf{P}_m, \tag{2}$$

where $\mathbf{F}_m^{token} \in \mathbb{R}^{C' \times \frac{DHW}{2^{3(l-1)}}}$ denotes the token and $\mathbf{W}_m$ denotes the weights of linear projection. The MSA builds the relationship within each modality by looking over all possible locations in the feature map, which is formulated as

$$head_m^i = Attention(\mathbf{Q}_m^i, \mathbf{K}_m^i, \mathbf{V}_m^i) = softmax(\frac{\mathbf{Q}_m^i \mathbf{K}_m^{iT}}{\sqrt{d_k}})\mathbf{V}_m^i, \tag{3}$$

$$MSA_m = [head_m^1, ..., head_m^N]\mathbf{W}_m^o, \tag{4}$$

where $\mathbf{Q}_m^i = LN(\mathbf{F}_m^{token})\mathbf{W}_m^{Qi}$, $\mathbf{K}_m^i = LN(\mathbf{F}_m^{token})\mathbf{W}_m^{Ki}$, $\mathbf{V}_m^i = LN(\mathbf{F}_m^{token})\mathbf{W}_m^{Vi}$, $LN(\cdot)$ is layer normalization, d_k is the dimension of $\mathbf{K}_m$, $N = 8$ is the number of attention heads, and $[\cdot, \cdot]$ is a concatenation operation. The FFN is a two-layer perceptron with GELU [9] activation. The feature maps with global context within each modality produced by the intra-modal Transformer is defined as

$$\mathbf{F}_m^{global} = FFN_m(LN(z)) + z, z = MSA_m(LN(\mathbf{F}_m^{token})) + \mathbf{F}_m^{token}, \tag{5}$$

where $\mathbf{F}_m^{global} \in \mathbb{R}^{C' \times \frac{DHW}{2^{3(l-1)}}}$.

2.2   Modality-Correlated Encoder

The modality-correlated encoder is designed to build the long-range correlations across modalities for modality-invariant features with global semantics corresponding to the tumor region. It is implemented as an inter-modal Transformer.

Inter-modal Transformer. In contrast to the intra-modal Transformer, the inter-modal Transformer combines the embeddings from all modality-specific encoders by concatenation as the input multimodal token, which is defined as

$$\mathbf{F}^{token} = [\delta_{FLAIR}\mathbf{F}_{FLAIR}^{global}, \delta_{T1c}\mathbf{F}_{T1c}^{global}, \delta_{T1}\mathbf{F}_{T1}^{global}, \delta_{T2}\mathbf{F}_{T2}^{global}]\mathbf{W} + \mathbf{P}, \tag{6}$$

where $\delta_m \in \{0, 1\}$ is a Bernoulli indicator that aims to grant robustness when building long-range dependencies between different modalities even when some modalities are missing. This kind of modality-level dropout is randomly conducted during training by setting δ_m to 0. In case of missing modalities, the multimodal token for the missing modalities will be held by a zero vector. Subsequently, it is processed by MSD and FFN for modality-invariant features across modalities, which is formulated as

$$\mathbf{F}^{global} = FFN(LN(z)) + z, z = MSA(LN(\mathbf{F}^{token})) + \mathbf{F}^{token}, \qquad (7)$$

where $\mathbf{F}^{global} \in \mathbb{R}^{C' \times \frac{DHW}{2^{(l-1)}}}$.

2.3   Convolutional Decoder

The convolutional decoder is designed to progressively restore the spatial resolution from high-level latent space to original mask space. The output sequence $\mathbf{F}^{global}$ of the modality-correlated Transformer is reshaped into feature maps corresponding to the size before flattening. The convolutional decoder has a symmetric architecture of convolutional encoder, similar to U-Net [15]. Besides, the skip connections between encoder and decoder are also added to keep more low-level details for better segmentation. The features from convolutional encoders of different modalities at a specific level are concatenated and forwarded as skip features to the convolutional decoder.

2.4   Auxiliary Regularizer

Conventional multimodal learning models tend to recognize brain tumors relying on the discriminative modalities [1,3]. Such models are likely to face severe degradation when the discriminative modalities are missing. Therefore, it is critical to encourage each convolutional encoder to segment brain tumors even without the assistance of other modalities. To this end, the outputs of convolutional encoders are upsampled by a shared-weight decoder to segment tumors from each modality separately. The shared-weight decoder has the same architecture with the convolutional decoder. Besides, we also introduce auxiliary regularizers in the convolutional decoder to force the decoder to generate accurate segmentation even when certain modalities are missing. It is achieved by interpolating the feature maps in each stage of the convolutional decoder to segment tumors via deep supervision [6]. Dice loss [13] is employed as the regularizer. Combining the training loss of the network's output with the auxiliary regularizers, the overall loss function is defined as

$$\mathcal{L} = 1 - Dice = 1 - \frac{2 \sum_{c=1}^{C} \sum_{i=1}^{N_c} g_i^c p_i^c}{\sum_{c=1}^{C} \sum_{i=1}^{N_c} g_i^{c2} + \sum_{c=1}^{C} \sum_{i=1}^{N_c} p_i^{c2}}, \qquad (8)$$

$$\mathcal{L}_{\text{total}} = \sum_{i \in M} \mathcal{L}_i^{encoder} + \sum_{i=1}^{l-1} \mathcal{L}_i^{decoder} + \mathcal{L}^{output}, \qquad (9)$$

Table 1. Results of the proposed method and state-of-the-art unified models, i.e., HeMIS [8] and U-HVED [4], on BraTS 2018 dataset [12]. Dice similarity coefficient (DSC) [%] is employed for evaluation with every combination settings of modalities. ● and ○ denote available and missing modalities, respectively.

Modalities				Enhancing Tumor			Tumor Core			Whole Tumor		
F	T1c	T1	T2	U-HeMIS	U-HVED	Ours	U-HeMIS	U-HVED	Ours	U-HeMIS	U-HVED	Ours
●	○	○	○	11.78	23.80	**39.33**	26.06	57.90	**61.21**	52.48	84.39	**86.10**
○	●	○	○	62.02	57.64	**72.60**	65.29	59.59	**75.41**	61.53	53.62	**72.22**
○	○	●	○	10.16	8.60	**32.53**	37.39	33.90	**56.55**	57.62	49.51	**67.52**
○	○	○	●	25.63	22.82	**43.05**	57.20	54.67	**64.20**	80.96	79.83	**81.15**
●	●	○	○	66.10	68.36	**75.07**	71.49	75.07	**77.88**	68.99	85.93	**87.30**
●	○	●	○	10.71	27.96	**42.96**	41.12	61.14	**65.91**	64.62	85.71	**87.06**
●	○	○	●	30.22	32.31	**47.52**	57.68	62.70	**69.75**	82.95	87.58	**87.59**
○	●	●	○	66.22	61.11	**74.04**	72.46	67.55	**78.59**	68.47	64.22	**74.42**
○	●	○	●	67.83	67.83	**74.51**	76.64	73.92	**78.61**	82.48	81.32	**82.99**
○	○	●	●	32.39	24.29	**44.99**	60.92	56.26	**69.42**	**82.41**	81.56	82.20
●	●	●	○	68.54	68.60	**75.47**	76.01	77.05	**79.80**	72.31	86.72	**87.33**
●	●	○	●	68.72	68.93	**75.67**	77.53	76.75	**79.55**	83.85	88.09	**88.14**
●	○	●	●	31.07	32.34	**47.70**	60.32	63.14	**71.52**	83.43	**88.07**	87.75
○	●	●	●	69.92	67.75	**74.75**	78.96	75.28	**80.39**	**83.94**	82.32	82.71
●	●	●	●	70.24	69.03	**77.61**	79.48	77.71	**85.78**	84.74	88.46	**89.64**
Average				46.10	46.76	**59.85**	62.57	64.84	**72.97**	74.05	79.16	**82.94**

where C is the number of segmentation classes, and N_c is the number of voxels of class c, g_i^c is a binary indicator if class label c is the correct classification for pixel i, p_i^c is the corresponding predicted probability, $M = \{FLAIR, T1c, T1, T2\}$, and l is the number of stages in the convolutional decoder.

3   Experiments and Results

Dataset and Implementation. The experiments are conducted on BraTS 2018 dataset[1] [12], which consists of 285 multi-contrast MRI scans with four modalities: T1, T1c, T2, and FLAIR. Different subregions of brain tumors are combined into three nested subregions: whole tumor, tumor core, and enhancing tumor. All the volumes have been co-registered to the same anatomical template and interpolated to the same resolution by the organizers. Dice Similarity Coefficient (DSC) as defined in Eq. (8) is employed for evaluation. The framework is implemented with PyTorch 1.7 on four NVIDIA Tesla V100 GPUs. The input size is $128 \times 128 \times 128$ voxels and batch size is 1. Random flip, crop, and intensity shifts are employed for data augmentation. The mmFormer has 106M parameters and 748G FLOPs. The network is trained with the Adam optimizer with an initial learning rate of 0.0002 for 1000 epochs. The model is trained for about 25 h with 17G memory on each GPU.

---

[1] https://www.med.upenn.edu/sbia/brats2018/data.html.

Performance of Incomplete Multimodal Segmentation. We evaluate the robustness of our method to incomplete multimodal segmentation. The absence of modality is implemented by setting $\delta_i, i \in \{FLAIR, T1c, T1, T2\}$ to be zero for dropping the specific modalities at inference. We compare our method with two representative models using shared latent space, i.e., HeMIS [8] and U-HVED [4]. For a fair comparison, we use the same data split in [21] and directly reference the results. In Table 1, our method significantly outperforms HeMIS and U-HVED on the segmentation of enhancing tumor and tumor core on all the 15 possible combinantions of available modalities and the segmentation of the whole tumor on 12 out of 15. In Table 2, we show that with the increased number of missing modalities, the average improvement obtained by mmFormer is more considerable. Meanwhile, it is observed that mmFormer gains more improvement when the target is more difficult to segment. These results demonstrate the effectiveness of mmFormer for incomplete multimodal learning of brain tumor segmentation. Figure 2 shows that even with one modality available, mmFormer can achieve proper segmentation for brain tumor.
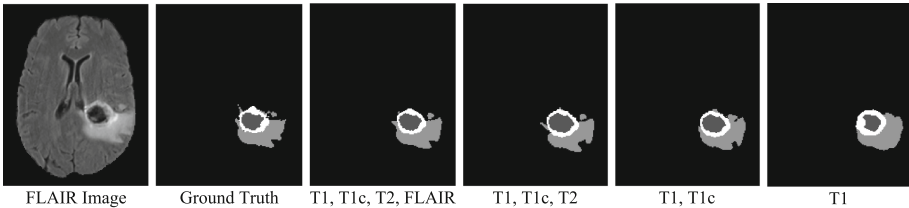

| FLAIR Image | Ground Truth | T1, T1c, T2, FLAIR | T1, T1c, T2 | T1, T1c | T1 |

Fig. 2. Segmentation results of mmFormer with various available modalities.

We also compare mmFormer with ACN [21]. ACN relies on knowledge distillation for incomplete multimodal brain tumor segmentation. In the case of N modalities in total, ACN has to train $2^4 - 2$ times to distill $2^N - 2$ student models for all conditions of missing modalities, while our mmFormer only learns once by a unified model. Specifically, ACN is trained for 672 h with 144M parameters for 1 teacher and 14 student models, while mmFormer requires only 25 h with 106 M parameters. Nevertheless, the average DSC for enhancing tumor, tumor core, and whole tumor of mmFormer (59.85, 72.97 and 82.94, respectively) is still close to it of ACN (61.21, 77.62, and 85.92, respectively).

Performance of Complete Multimodal Segmentation. We compare our method with a recent Transformer-based method, i.e., TransBTS [20], for multimodal brain tumor segmentation with full modalities. We reproduce the results with the official repository. TransBTS obtains DSC of 72.66%, 72.69%, and 79.99% on enhancing tumor, tumor core, and the whole tumor, respectively. Our mmFormer outperforms TransBTS on all subregions of brain tumor with DSC of 77.61%, 85.78%, and 89.64%, demonstrating the effectiveness of mmFormer even for complete multimodal brain tumor segmentation.

Ablation Study. We investigate the effectiveness of intra-modal Transformer, inter-modal Transformer, and auxiliary regularizer as three critical components in our method. We analyze the effectiveness of each component by excluding one of them from mmFormer. In Table 3, we compare the performance of the three variants to mmFormer with DSC, averaging over the 15 possible combinations of input modalities. It shows that intra-modal Transformer, inter-modal Transformer, and auxiliary regularizer bring performance improvement across all the tumor subregions.

Table 2. Average improvements of mmFormer upon HeMIS [8] and U-HVED [4] with different numbers of missing modalities evaluated by DSC [%].

Regions	# of missing modalities			
	0	1	2	3
Enhancing	+7.98	+8.91	+13.57	+19.07
Core	+7.19	+4.68	+8.62	+15.34
Whole	+3.04	+2.89	+5.57	+11.75

Table 3. Ablation study of critical components of mmFormer.

Methods	Average DSC [%]		
	Enhancing	Core	Whole
mmFormer	59.85	72.97	82.94
w/o IntraTrans	56.98	71.83	81.32
w/o InterTrans	56.05	70.28	81.12
w/o Aux. Reg	55.78	69.33	81.65

4   Conclusion

We proposed a Transformer-based method for incomplete multimodal learning of brain tumor segmentation. The proposed mmFormer bridges Transformer and CNN to build the long-range dependencies both within and across different modalities of MRI images for a modality-invariant representation. We validated our method on brain tumor segmentation under various combinations of missing modalities, and it outperformed state-of-the-art methods on the BraTS benchmark. Our method gains more improvements when more modalities are missing and/or the target ones are more difficult to segment.

References

1. 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: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11766, pp. 447–456. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_50
2. Chen, C., Dou, Q., Jin, Y., Liu, Q., Heng, P.A.: Learning with privileged multimodal knowledge for unimodal segmentation. IEEE Trans. Med. Imaging **41**, 621–632 (2021)
3. 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)

4. Dorent, R., Joutard, S., Modat, M., Ourselin, S., Vercauteren, T.: Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 74–82. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_9

5. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

6. Dou, Q., et al.: 3D deeply supervised network for automated segmentation of volumetric medical images. Med. Image Anal. **41**, 40–54 (2017)

7. Hatamizadeh, A., et al.: UNETR: transformers for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584 (2022)

8. Havaei, M., Guizard, N., Chapados, N., Bengio, Y.: HeMIS: hetero-modal image segmentation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 469–477. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_54

9. Hendrycks, D., Gimpel, K.: Gaussian error linear units (GELUs). arXiv preprint arXiv:1606.08415 (2016)

10. Hu, M., et al.: Knowledge distillation from multi-modal to mono-modal segmentation networks. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 772–781. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_75

11. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)

12. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2014)

13. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: Fourth International Conference on 3D Vision, pp. 565–571. IEEE (2016)

14. Peiris, H., Hayat, M., Chen, Z., Egan, G., Harandi, M.: A volumetric transformer for accurate 3D tumor segmentation. arXiv preprint arXiv:2111.13300 (2021)

15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

16. Shen, Y., Gao, M.: Brain tumor segmentation on MRI with missing modalities. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (eds.) IPMI 2019. LNCS, vol. 11492, pp. 417–428. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20351-1_32

17. Tseng, K.L., Lin, Y.L., Hsu, W., Huang, C.Y.: Joint sequence learning and cross-modality convolution for 3D biomedical segmentation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 6393–6400 (2017)

18. van Tulder, G., de Bruijne, M.: Why does synthesized data improve multi-sequence classification? In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 531–538. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24553-9_65

19. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30 (2017)

20. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: TransBTS: multimodal brain tumor segmentation using transformer. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 109–119. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_11

21. Wang, Y., et al.: ACN: adversarial co-training network for brain tumor segmentation with missing modalities. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12907, pp. 410–420. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87234-2_39

22. Zhang, Y., et al.: Modality-aware mutual learning for multi-modal medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 589–599. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_56

23. Zhou, C., Ding, C., Lu, Z., Wang, X., Tao, D.: One-pass multi-task convolutional neural networks for efficient brain tumor segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11072, pp. 637–645. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00931-1_73