A2FSeg: Adaptive Multi-modal Fusion Network for Medical Image Segmentation

Zirui Wang and Yi Hong(✉)

Department of Computer Science and Engineering, Shanghai Jiao Tong University,
Shanghai 200240, China
yi.hong@sjtu.edu.cn

Abstract. Magnetic Resonance Imaging (MRI) plays an important role in multi-modal brain tumor segmentation. However, missing modality is very common in clinical diagnosis, which will lead to severe segmentation performance degradation. In this paper, we propose a *simple* adaptive multi-modal fusion network for brain tumor segmentation, which has two stages of feature fusion, including a simple average fusion and an adaptive fusion based on an attention mechanism. Both fusion techniques are capable to handle the missing modality situation and contribute to the improvement of segmentation results, especially the adaptive one. We evaluate our method on the BraTS2020 dataset, achieving the state-of-the-art performance for the incomplete multi-modal brain tumor segmentation, compared to four recent methods. Our A2FSeg (Average and Adaptive Fusion Segmentation network) is simple yet effective and has the capability of handling any number of image modalities for incomplete multi-modal segmentation. Our source code is online and available at https://github.com/Zirui0623/A2FSeg.git.

Keywords: Modality-adaptive fusion · Missing modality · Brain tumor segmentation · Incomplete multi-modal segmentation

1 Introduction

Extracting brain tumors from medical image scans plays an important role in further analysis and clinical diagnosis. Typically, a brain tumor includes peritumoral edema, enhancing tumor, and non-enhancing tumor core. Since different modalities present different clarity of brain tumor components, we often use multi-modal image scans, such as T1, T1c, T2, and Flair, in the task of brain tumor segmentation [12]. Works have been done to handle brain tumor segmentation using image scans collected from all four modalities [11,15]. However, in practice, we face the challenge of collecting all modalities at the same time, with often one or more missing. Therefore, in this paper, we consider the problem of segmenting brain tumors with missing image modalities.

Current image segmentation methods for handling missing modalities can be divided into three categories, including: 1) brute-force methods: designing individual segmentation networks for each possible modality combination [18], 2)

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14223, pp. 673–681, 2023.
https://doi.org/10.1007/978-3-031-43901-8_64

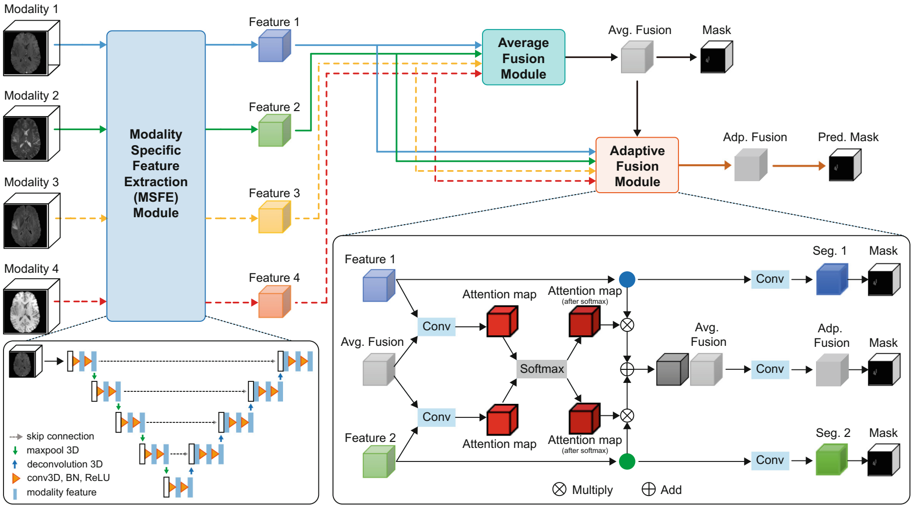

Fig. 1. Overview of our proposed adaptive multi-modal fusion network (A2FSeg, short for Average and Adaptive Fusion Segmentation network). The dashed lines indicate the possibility of missing some modalities. If so, both the average fusion module and the adaptive fusion module will ignore the missing ones. The final tumor mask is predicted based on feature maps after the adaptive fusion, indicated by the solid red arrows. (Best viewed in color) (Color figure online)

completion methods: synthesizing the missing modalities to complete all modalities required for conventional image segmentation methods [16], and 3) fusion-based methods: mapping images from different modalities into the same feature space for fusion and then segmenting brain tumors based on the fused features [10]. Methods in the first category have good segmentation performance; however, they are resource intensive and often require more training time. The performance of methods in the second category is limited by the synthesis quality of the missing modality. The third category often has one single network to take care of different scenarios of missing modalities, which is the most commonly used one in practice.

To handle various numbers of modal inputs, HeMIS [5] projects the image features of different modalities into the same feature space, by computing the mean and variance of the feature maps extracted from different modalities as the fused features. To improve the representation of feature fusion, HVED [3] treats the input of each modality as a Gaussian distribution, and fuses feature maps from different modalities through a Gaussian mixture model. RobustSeg [1], on the other hand, decomposes the modality features into modality-invariant content code and modality-specific appearance code, for more accurate fusion and segmentation. Considering the different clarity of brain tumor regions observed in different modalities, RFNet [2] introduces an attention mechanism to model the relations of modalities and tumor regions adaptively. Based on graph structure and attention mechanism, MFI [21] is proposed to learn adaptive complementary information between modalities in different missing situations.

Due to the complexity of current models, we tend to develop a simple model, which adopts a simple average fusion and attention mechanism. These two techniques are demonstrated to be effective in handling missing modalities and multimodal fusion [17]. Inspired by MAML [20], we propose a model called A2FSeg (Average and Adaptive Fusion Segmentation network, see Fig. 1), which has two fusion steps, i.e., an average fusion and an attention-based adaptive fusion, to integrate features from different modalities for segmentation. Although our fusion idea is quite simple, A2FSeg achieves state-of-the-art (SOTA) performance in the incomplete multimodal brain tumor image segmentation task on the BraTS2020 dataset. Our contributions in this paper are summarized below:

- We propose a *simple* multi-modal fusion network, A2FSeg, for brain tumor segmentation, which is general and can be extended to any number of modalities for incomplete image segmentation.
- We conduct experiments on the BraTS 2020 dataset and achieve the SOTA segmentation performance, having a mean Dice core of 89.79% for the whole tumor, 82.72% for the tumor core, and 66.71% for the enhancing tumor.

2 Method

Figure 1 presents the network architecture of our A2FSeg. It consists of four modality-specific sub-networks to extract features from each modality, an average fusion module to simply fuse features from available modalities at the first stage, and an adaptive fusion module based on an attention mechanism to adaptively fuse those features again at the second stage.

Modality-Specific Feature Extraction (MSFE) Module. Before fusion, we first extract features for every single modality, using the nnUNet model [7] as shown in Fig. 1. In particular, this MSFE model takes a 3D image scan from a specific modality m, i.e., $\mathbf{I}_m \in \mathbb{R}^{H \times W \times D}$ and $m \in \{\text{T1, T2, T1c, Flair}\}$, and outputs the corresponding image features $\mathbf{F}_m \in \mathbb{R}^{C \times H_f \times W_f \times D_f}$. Here, the number of channels is $C = 32$; H_f, W_f, and D_f are the height, width, and depth of feature maps $\mathbf{F}_m$, which share the same size as the input image. For every single modality, each MSFE module is supervised by the image segmentation mask to fasten its convergence and provide a good feature extraction for fusion later. All four MSFEs have the same architecture but with different weights.

Average Fusion Module. To aggregate image features from different modalities and handle the possibility of missing one or more modalities, we use the average of the available features from different modalities as the first fusion result. That is, we obtain a fused average feature $\bar{\mathbf{F}} = \frac{1}{N_m} \sum_{m=1}^{N_m} \mathbf{F}_m$. Here, N_m is the number of available modalities. For example, as shown in Fig. 1, if only the first two modalities are available at an iteration, then $N_m = 2$, and we will take the average of these two modalities, ignoring those missing ones.

Adaptive Fusion Module. Since each modality contributes differently to the final tumor segmentation, similar to MAML [20], we adopt the attention mechanism to measure the voxel-level contributions of each modality to the final segmentation. As shown in Fig. 1, to generate the attention map for a specific modality m, we take the concatenation of its feature extracted by the MSFE module $\mathbf{F}_m$ and the mean feature after the average fusion $\bar{\mathbf{F}}$, which is passed through a convolutional layer to generate the initial attention weights:

$$\mathbf{W}_m = \sigma\left(\mathcal{F}_m\left(\left[\bar{\mathbf{F}}; \mathbf{F}_m\right]; \theta_m\right)\right), \quad m \in \{\text{T1, T1c, T2, Flair}\}. \tag{1}$$

Here, $\mathcal{F}_m$ is a convolutional layer for this specific modality m, and θ_m represents the parameters of this layer, and σ is a Sigmoid function. That is, we have an individual convolution layer $\mathcal{F}_m$ for each modality to generate different weights.

Due to the possibility of missing modalities, we will have different numbers of feature maps for fusion. To address this issue, we normalize the different attention weights by using a Softmax function:

$$\hat{\mathbf{W}}_m = \frac{\exp\left(\mathbf{W}_m\right)}{\sum_m^{N_m} \exp\left(\mathbf{W}_m\right)}. \tag{2}$$

That is, we only consider feature maps from those available modalities but normalize their contribution to the final fusion result, so that, the fused one has a consistent value range, no matter how many modalities are missing. Then, we perform voxel-wise multiplication of the attention weight with the corresponding modal feature maps. As a result, the adaptively fused feature maps $\hat{\mathbf{F}}$ is calculated by the weighted sum of each modal feature:

$$\hat{\mathbf{F}} = \sum_m \hat{\mathbf{W}}_m \otimes \mathbf{F}_m. \tag{3}$$

Here, $\otimes$ indicates the voxel-wise multiplication.

Loss Function. We have multiple segmentation heads, which are distributed in each module of A2FSeg. For each segmentation head, we use the combination of the cross-entropy and the soft dice score as the basic loss function, which is defined as

$$\mathcal{L}(\hat{y}, y) = \mathcal{L}_{CE}(\hat{y}, y) + \mathcal{L}_{Dice}(\hat{y}, y), \tag{4}$$

where $\hat{y}$ and y represent the segmentation prediction and the ground truth, respectively. Based on this basic one, we have the overall loss function defined as

$$\mathcal{L}_{total} = \sum_m \mathcal{L}_m(\hat{y}_m, y) + \mathcal{L}_{avg}(\hat{y}_{avg}, y) + \mathcal{L}_{adp}(\hat{y}_{adp}, y), \tag{5}$$

where the first term is the basic segmentation loss for each modality m after feature extraction; the second term is the loss for the segmentation output of the average fusion module; and the last term is the segmentation loss for the final output from the adaptive fusion module.

Table 1. Comparison among recent methods, including HeMIS [5], U-HVED [3], mmFormer [19], and MFI [21], and ours on BraTS2020 in terms of Dice%. Missing and available modalities are denoted by ○ and ●, respectively. F indicates Flair, HVED indicates U-HVED, and Former indicates mmFormer because of space issue.

Modalities				Complete					Core					Enhancing				
T_1	T_{1c}	T_2	F	Hemis	HVED	Former	MFI	Ours	Hemis	HVED	Former	MFI	Ours	Hemis	HVED	Former	MFI	Ours
○	○	○	●	87.76	86.49	90.08	90.60	**91.48**	66.56	64.42	71.13	75.59	**76.21**	44.95	43.32	48.25	51.96	**53.80**
○	○	●	○	85.53	85.14	87.00	88.38	**88.82**	65.55	64.87	72.85	75.38	**76.40**	43.77	43.31	50.18	52.72	**54.46**
○	○	●	●	90.51	89.87	91.19	91.65	**91.95**	70.82	70.55	75.18	77.42	**77.83**	48.32	47.86	52.51	54.77	**56.10**
○	●	○	○	72.83	74.31	80.00	80.16	**83.11**	83.59	83.96	85.29	85.35	**86.95**	75.54	77.34	76.17	76.91	**78.01**
○	●	○	●	91.29	90.45	91.51	92.36	**92.42**	86.27	85.78	87.05	**88.67**	87.96	76.30	76.29	76.99	77.26	**78.07**
○	●	●	○	86.32	86.82	88.79	89.53	**89.90**	85.61	85.11	87.41	**87.83**	87.75	75.57	75.68	77.46	76.56	**77.85**
○	●	●	●	91.82	91.46	91.93	92.38	**92.72**	86.63	86.06	87.87	**88.56**	87.96	76.25	75.47	76.15	76.69	**76.96**
●	○	○	○	75.02	76.64	81.20	79.91	**83.67**	61.18	62.78	71.36	72.36	**75.52**	37.55	39.46	46.65	50.40	**52.58**
●	○	○	●	90.29	88.81	91.29	91.45	**91.89**	71.95	70.18	76.01	**78.22**	78.07	48.16	46.53	51.20	**55.05**	54.00
●	○	●	○	86.66	87.13	88.22	88.03	**89.40**	67.67	70.21	75.00	75.85	**77.39**	44.86	46.95	51.37	54.39	**54.58**
●	○	●	●	90.85	90.34	91.61	91.67	**92.23**	72.75	73.22	77.05	78.30	**78.64**	48.48	49.45	52.51	**55.44**	55.34
●	●	○	○	77.42	79.40	82.53	82.50	**84.81**	84.76	84.94	86.03	86.52	**87.40**	75.43	76.56	76.84	76.76	**77.80**
●	●	○	●	91.65	90.97	91.95	92.24	**92.29**	86.79	86.61	87.44	**88.84**	87.75	76.44	75.79	76.91	**77.06**	76.75
●	●	●	○	86.75	87.72	89.19	88.81	**89.49**	86.11	85.36	**87.30**	87.22	87.16	75.16	75.62	76.37	76.52	**77.69**
●	●	●	●	92.00	91.62	92.26	92.33	**92.71**	87.67	86.46	88.13	**88.60**	87.74	75.39	75.66	76.08	76.66	**76.70**
Means				86.45	86.48	88.58	88.80	**89.79**	77.59	77.37	81.01	82.31	**82.72**	61.48	61.69	64.38	65.94	**66.71**

3 Experiments

3.1 Dataset

Our experiments are conducted on BraTS2020, which contains 369 multi-contrast MRI scans with four modalities: T1, T1c, T2, and Flair. These images went through a sequence of preprocessing steps, including co-registration to the same anatomical template, resampling to the same resolution ($1\,\text{mm}^3$), and skull-stripping. The segmentation masks have three labels, including the whole tumor (abbreviated as Complete), tumor core (abbreviated as Core), and enhancing tumor (abbreviated as Enhancing). These annotations are manually provided by one to four radiologists according to the same annotation protocol.

3.2 Experimental Settings and Implementation Details

We implement our model with PyTorch [13] and perform experiments on an Nvidia RTX3090 GPU. We use the Adam optimizer [8], with an initial learning rate of 0.01. Since we use the method of exponential decay of learning rate, the initial learning rate is then multiplied by $(1 - \frac{\#\text{epoch}}{\#\text{max\_epoch}})^{0.9}$. Due to the limitation of GPU memory, each volume is randomly cropped into multiple patches with the size of $128 \times 128 \times 128$ for training. The network is trained for 400 epochs. In the inference stage, we use a sliding window to produce the final segmentation prediction of the input image.

3.3 Experimental Results and Comparison to Baseline Methods

To evaluate the performance of our model, we compare it with four recent models, HeMIS [5], U-HVED [3], mmFormer [19], and MFI [21]. The dataset is randomly split into 70% for training, 10% for validation, and 20% for testing, and all methods are evaluated on the same dataset and data splitting. We use the Dice score

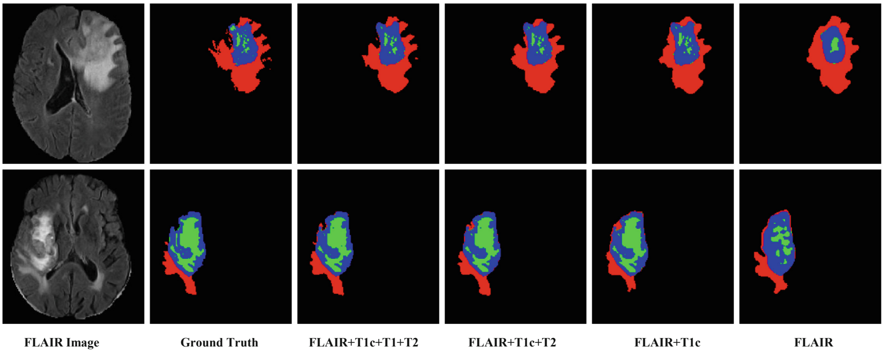

Fig. 2. Visualization of our A2FSeg results using a different number of modalities for brain tumor segmentation. Red: peritumoral edema; Blue: enhancing tumor; Green: the necrotic and non-enhancing tumor core. (Color figure online)

as the metric. As shown in Table 1, our method achieves the best result. For example, our method outperforms the current SOTA method MFI [21] in most missing-modality cases, including all cases for the whole/complete tumor, 8 out of 15 cases for the tumor core, 12 out of 15 cases for the enhancing tumor. Compared to MFI, for the whole tumor, tumor core, and enhancing tumor regions, we improve the average Dice scores by 0.99%, 0.41%, and 0.77%, respectively. Although the design of our model is quite simple, these results demonstrate its effectiveness for the incomplete multimodel segmentation task of brain tumors.

Figure 2 visualizes the segmentation results of samples from the BraTS2020 dataset. With only one Flair image available, the segmentation results of the tumor core and enhancing tumor are poor, because little information on these two regions is observed in the Flair image. With an additional T1c image, the segmentation results of these two regions are significantly improved and quite close to the ground truth. Although adding T1 and T2 images does not greatly improve the segmentation of the tumor core and the enhancing tumor, the boundary of the whole tumor is refined with their help.

Figure 3 visualizes the contribution to each tumor region from each modality. The numbers are the mean values of the attention maps computed for images in the test set. Overall, in our model, each modality has its contribution to the final segmentation, and no one dominates the result. This is because we have supervision on the segmentation branch of each modality, so that, each modality has the ability to segment each region to some extent. However, we still observe that Flair and T2 modalities have relatively larger contributions to the segmentation of all tumor regions, followed by T1c and then T1. This is probably because the whole tumor area is much clear in Flair and T2 compared to the other two modalities. Each modality shows its preference when segmenting different regions. Flair and T2 are more useful for extracting the peritumoral edema (ED) than the enhancing tumor (ET) and the non-enhancing tumor and

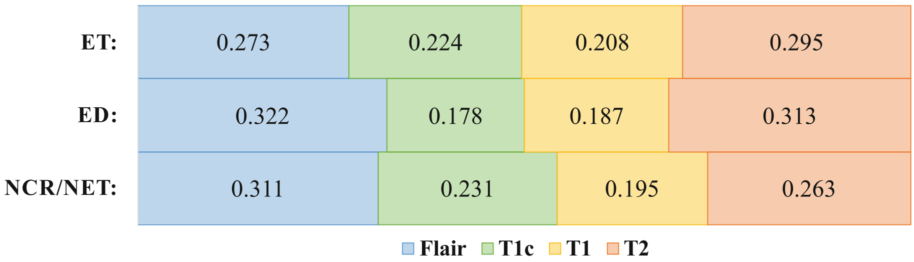

Fig. 3. Summary of the contribution of each modality to each tumor region from the estimated attention maps. ET: enhancing tumor, ED: the peritumoral edema, NCR/NET: non-enhancing tumor and necrosis.

Table 2. Ablation study of the adaptive fusion model in our method.

Methods	Complete	Core	Enhancing	Average
MFI	88.80	82.31	65.94	79.02
Baseline (Average fusion module only)	89.29	82.00	66.00	79.10
+Adaptive fusion module	**89.79**	**82.72**	**66.71**	**79.74**

necrosis (NCR/NET); while T1c and T1 are on the opposite and more helpful for extracting ET and NCR/NET.

3.4 Ablation Study

In this part, we investigate the effectiveness of the average fusion module and the adaptive fusion module, which are two important components of our method. Firstly, we set a baseline model without any modal interaction, that is, with the average fusion module only. Then, we add the adaptive fusion module to the baseline model. Table 2 reports this ablation study. With only adding the average fusion module, our method already obtains comparable performance with the current SOTA method MFI. By adding the adaptive fusion module, the dice scores of the three regions further increase by 0.50%, 0.72%, and 0.71%, respectively. This shows that both the average fusion module and the adaptive fusion module are effective in this brain tumor segmentation task.

4 Discussion and Conclusion

In this paper, we propose an average and adaptive fusion segmentation network (A2FSeg) for the incomplete multi-model brain tumor segmentation task. The essential components of our A2FSeg network are the two stages of feature fusion, including an average fusion and an adaptive fusion. Compare to existing complicated models, our model is much simpler and more effective, which is demonstrated by the best performance on the BraTS 2020 brain tumor segmentation task. The experimental results demonstrate the effectiveness of two techniques, i.e., the average fusion and the attention-based adaptive one, for incomplete modal segmentation tasks.

Our study brings up the question of whether having complicated models is necessary. If there is no huge gap between different modalities, like in our case where all four modalities are images, the image feature maps are similar and a simple fusion like ours can work. Otherwise, we perhaps need an adaptor or an alignment strategy to fuse different types of features, such as images and audio.

Also, we observe that a good feature extractor is essential for improving the segmentation results. In this paper, we only explore a reduced UNet for feature extraction. In future work, we will explore other feature extractors, such as Vision Transformer (ViT) or other pre-trained visual foundation models [4,6,14]. Recently, the segment anything model (SAM) [9] demonstrates its general ability to extract different regions of interest, which is promising to be adopted as a good starting point for brain tumor segmentation. Besides, our model is general for multi-modal segmentation and we will apply it to other multi-model segmentation tasks to evaluate its generalization on other applications.

Acknowledgements. This work was supported by NSFC 62203303 and Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102.

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, Part III. LNCS, vol. 11766, pp. 447–456. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_50
2. 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)
3. 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, Part II. LNCS, vol. 11765, pp. 74–82. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_9
4. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale (2021)
5. 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, Part II. LNCS, vol. 9901, pp. 469–477. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_54
6. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16000–16009 (2022)
7. 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. Nat. Methods **18**(2), 203–211 (2021)

8. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Kirillov, A., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
10. Li, R., et al.: Deep learning based imaging data completion for improved brain disease diagnosis. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014, Part III. LNCS, vol. 8675, pp. 305–312. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10443-0_39
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. Paszke, A., et al.:Automatic differentiation in pytorch (2017)
14. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, 18–24 July 2021, vol. 139, pp. 8748–8763. PMLR (2021), https://proceedings.mlr.press/v139/radford21a.html
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, Part III. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
16. 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, Part I. LNCS, vol. 9349, pp. 531–538. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24553-9_65
17. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
18. 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, Part VII. LNCS, vol. 12907, pp. 410–420. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87234-2_39
19. Zhang, Y., et al.: mmFormer: multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022, Part V. LNCS, vol. 13435, pp. 107–117. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_11
20. Zhang, Y., et al.: Modality-aware mutual learning for multi-modal medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021, Part I. LNCS, vol. 12901, pp. 589–599. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_56
21. Zhao, Z., Yang, H., Sun, J.: Modality-adaptive feature interaction for brain tumor segmentation with missing modalities. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022, Part V. LNCS, vol. 13435, pp. 183–192. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_18