Enhancing Encoder with Attention Gate for Multimodal Brain Tumor Segmentation

Yi Li[1], Zhirui Fang[1], Di Li[2], Xin Xie[1], and Yanqing Guo[1,⋆]

[1] Dalian University of Technology, China
[2] Dalian Municipal Central Hospital, China

Abstract. Magnetic Resonance Imaging (MRI) is widely applied to diagnose malignant brain tumors like glioblastoma (GBM). Recent deep network based brain tumor segmentation algorithms have facilitated automatic and accurate segmentation on MRI data, benefiting the clinical diagnosis with efficiency. However, existing methods most work on certain datasets but suffer from performance degradation when tested on unseen out-of-sample datasets. In this paper, we integrate the encoder-decoder network structure with attention gate and Variational Autoencoders (VAE) to achieve promising segmentation results across different situations. Considering there are four modalities in each brain MRI sample, an encoder based on 3D convolution is employed to capture the local correlation among both spatial and modal neighbors. Then the extracted volumetric feature maps are fed into a decoder, finally generating the segmentation results with attention gate module. To facilitate better segmentation, we further adopt VAE as an auxiliary decoder to improve the performance of the encoder.

Keywords: Brain Tumor Segmentation · Variational Autoencoders · Attention Gate · Federated Evaluation · FeTS Challenge.

1 Introduction

Glioblastomas (GBM) are deemed as the most aggressive and heterogeneous adult brain tumor [16], with the median survival of approximately 15 months [15]. In practice, magnetic resonance imaging (MRI) offers an applicable choice for routine clinical diagnosis in GBM. There are usually four modalities in each MRI sample, including T1-weighted (T1), contrast-enhanced T1-weighted (T1c), T2-weighted (T2), and Fluid Attenuated Inversion Recovery (FLAIR) images. Since these modalities provides different pathology clues, it is of great importance to learn them comprehensively for better segmentation performance.

1.1 Medical Image Segmentation

With the rise of deep learning, Convolutional Neural Networks (CNN) based approaches have achieved remarkable progress in medial image segmentation.

[⋆] Corresponding author: guoqy@dlut.edu.cn

Among them, the Fully Convolutional Network (FCN) [12] is an epochal work that produces impressive segmentation results with an end-to-end network. It afterward is usually used as the feature extractor for medical image analysis. Another representative architecture in medical image segmentation is U-Net [18] which builds connections between the encoder layer and the corresponding decoder layer via feature map duplication. With these connections that skip the network bottleneck, lower level details are sent to the decoder for delicate segmentation outputs. Later literatures [8, 28, 14] continue to improve the U-Net architecture from different points of view. However, these methods are inevitably limited by the inductive locality bias of convolution, the reason coming from the marginal scale of the receptive field. Therefore, how to model the long-range dependencies becomes one of the breakthroughs in medical image segmentation.

1.2 Self-attention

Arising from natural language processing, the attention mechanism helps networks to capture long-range dependencies in feature maps. Many works [24, 19] have explored to combine the advantages of CNN and the attention mechanism. Recently, the transformer framework [22] is proposed and achieves the fantastic performance on sequence-to-sequence translation. The essence of the transformer is multiple self-attention layers, which can capture interactions between all pairs of elements in the input sequence regardless of their relative position. Now the transformer is also applied to computer vision tasks successfully. For example, it is introduced to image classification [9, 7, 1], 3-Dimensional video grounding [26, 21, 20], object detection [6, 27] and style Transfer[25, 11]. Despite the excellent and convincing results, the computational complexity of the transformer based approaches increase exponentially. The issue becomes even more serious in medical image analysis, because the qualified data can be very scarce for uncommon diseases like GBM. Therefore, how to balance the parameter scale of the transformer and the training data is an important problem to be solved.

1.3 the Generalization Problem

Although the approaches based on neural networks have witnessed great progresses in medical image segmentation, they still face challenges in practical scenarios, including “AI chasm”. “AI chasm” refers to the performance discrepancy of an AI algorithm in research environments and real-life applications. Algorithms based on networks are essentially data-driven and tend to be limited by the diversity of the training data. Existing methods are usually trained and tested on the subset of a dataset, sliding over the data discrepancy in practice. When evaluated on unseen out-of-sample datasets from various institutions that did not contribute data on the training set like the FeTS Challenge does, most deep learning models will experience performance deterioration. To measure the generalization ability, in the FeTS Challenge, the segmentation models are evaluated across different medical institutions, MRI scanners, image acquisition parameters and populations. Therefore, it has practical significance to tackle the distribution shift between the training and the test sets and thus raise the generalization ability of the model.

1.4 Method Motivation

On the basis of the above considerations, we summarize that an advanced method for multimodal brain tumor segmentation should have the following characters.

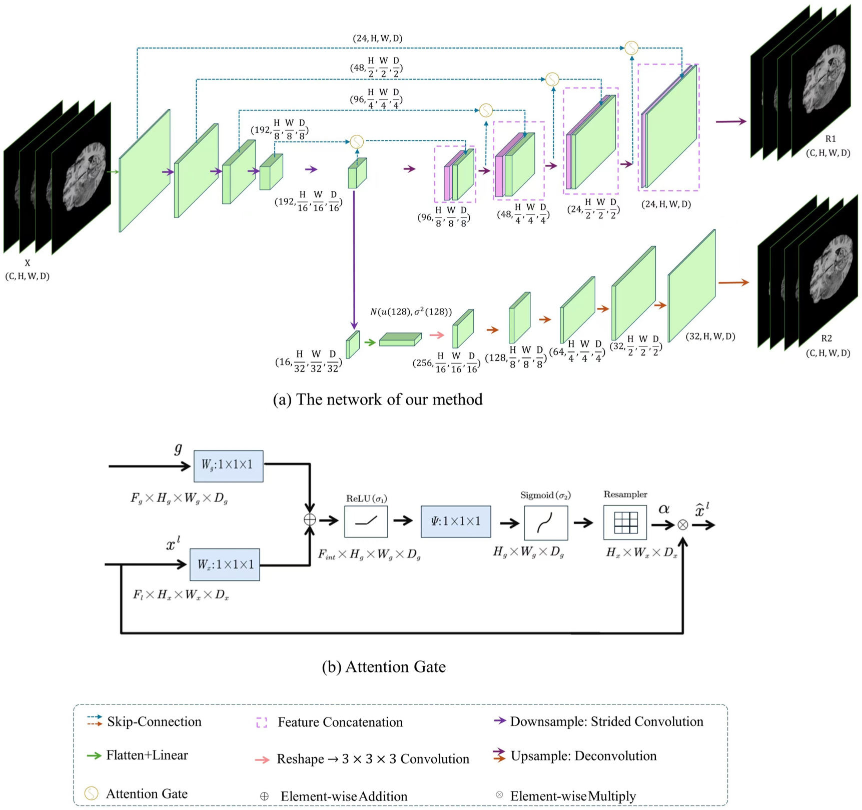
- Taking both encoded and decoded information in the medical image into account
- Exploiting the effective collaboration among different modalities of MRI
- Tackling the distribution discrepancy between the training and the test sets
- Producing accurate segmentation results with affordable computation cost

Inspired by the recent progress [23, 10, 2, 13, 4, 3, 5] in multimodal brain tumor segmentation, we implement a typical encoder-decoder structure. As in [23], instead of using 2D convolution to process the MRI sample slice-by-slice, a 3D CNN is employed to learn different modalities of MRI as a whole, which can capture the local features within as well as across MRI modalities. Different from [23], we further utilize an auxiliary VAE to enhance the ability of encoder in feature extraction. Besides, attention gates are used to conduct the skip connections for obtaining more accurate segmentation, which mitigates the over-fitting risk and benefits the model generalization ability.

2 Method

2.1 Overview

Fig. 1 (a) presents an illustration of the designed network consisting of roughly four components. It essentially follows the encoder-decoder structure, whereas the 3D CNN builds up the enhanced encoder together and there are two branches of decoder during training. Given an MRI sample $I \in \mathbb{R}^{C\times H\times W\times D}$, the 3D CNN first embeds the input into a feature map F, to capture the local knowledge within and across different modalities. Specifically, C refers to the number of modalities, $H \times W$ is the spatial resolution of the medical image, and D is the depth (or number of slices) of the medial image. Then the segmentation result is output by the chief decoder (the upper decoder in Fig. 1) (a) with a series of deconvolution layers. The auxiliary VAE (the lower decoder in Fig. 1) (a) is employed to help with the parameter learning of the 3D CNN during training. Following the U-Net structure, there are also skip connections through attention gates between the corresponding layers in the encoder and the decoder.



(a) The network of our method

(b) Attention Gate

Fig. 1. Overview of the designed network.

2.2 3D CNN

For the encoder component, we employ residual blocks, with each individual block comprising two convolutional layers accompanied by normalization and Rectified Linear Unit (ReLU) activation. Following this, an additive identity skip connection is incorporated. To achieve this, we utilize convolution operations employing a kernel size of $3 \times 3 \times 3$. This step allows for a gradual reduction in the dimensions of the image by a factor of 2, progressively integrating the nearby context into a feature map denoted as F within the real-number space $\mathbb{R}^{192\times\frac{H}{16}\times\frac{W}{16}\times\frac{D}{16}}$. Moreover, we apply Batch Normalization (BatchNorm) as the chosen normalization technique, which in turn leads to enhanced performance outcomes. By subjecting the input data to the 3D Convolutional Neural Network (CNN), we not only acquire more intricate local details, but we also alleviate the computational burden. This is due to the fact that the necessity to individually process each component or modality is obviated.

2.3 Attention Gate

Like [14], we utilize the attention gates (AGs) to alleviate feature loss through skip connections. The attention gate block is illustrated in Fig. 1 (b). The input features (x^l) undergo scaling using attention coefficients (α) computed within the AG module. Spatial regions are chosen by analyzing both the activations and contextual information derived from the gating signal (g), which is acquired from a more coarse-grained level. Resampling of attention coefficients on a grid is accomplished through trilinear interpolation. In this paper, we adopt multi-dimensional AGs where we extract and blend complementary information to establish the output of the skip connection. To alleviate the burden of excessive trainable parameters and the computational intricacy associated with AGs, we execute linear transformations without involving spatial support (utilizing $1 \times 1 \times 1$ convolutions). Moreover, we downsample the input feature-maps to match the resolution of the gating signal. This strategic approach ensures that attention units across various scales possess the capacity to influence responses encompassing a broad spectrum of foreground content within the images. Consequently, we proactively prevent the reconstruction of dense predictions solely from minute subsets of skip connections.

2.4 Decoder

The segmentation result $R_1 \in \mathbb{R}^{C\times H\times W\times D}$ is produced by the decoder with the intermediate feature F as the input. There are two decoders in the network shown in Fig. 1 (a), the upper being the chief decoder while the lower being the auxiliary decoder. The chief decoder is the U-Net architecture with the skip connections and holds individual parameters. Different from the chief decoder, the auxiliary decoder is the VAE network, sampling from the Gaussian distribution $N(\mu(128), \sigma^2(128))$. Hence the function of them differs from each other, the chief decoder aims at better segmentation results while the auxiliary one is to prevent the latent feature loss, which means encoder can better capture the tumor information. For the loss functions, we use the softmax Dice loss which can be calculated by

$$\mathcal{L}_{dice} = 1 - Dice(g_j^c, p_j^c) = 1 - \frac{2\sum_{c=1}^{M}\sum_{j=1}^{N_c} g_j^c p_j^c}{\sum_{c=1}^{M}\sum_{j=1}^{N_c}(g_j^c)^2 + \sum_{c=1}^{M}\sum_{j=1}^{N_c}(p_j^c)^2} \tag{1}$$

where g_j^c is a binary variable that indicates whether c is the correct label for position j, p_j^c is the predicted probability of label c at position j, M refers to the number of labels, and N_c represents the voxel number of label c in the sample. Since there are usually three types of regions to be concerned (including enhancing tumor region (ET), tumor core region (TC), and the whole tumor region (WT)), the total loss varies in the range of $[0, 3]$ theoretically.

Besides, we use the typical cross entropy loss to further promise the segmentation accuracy:

Table 1. The network details of our method.

Component	Block	Operation	Output size
Encoder: 3D CNN	InitConv	$\begin{bmatrix} Conv3, BN, ReLU \\ Conv3, BN, ReLU \end{bmatrix}$	$24 \times 160 \times 160 \times 128$
	DownSample i EncBlock i (i = 1, 2, 3)	Maxpool(kernel2) $\begin{bmatrix} Conv3, BN, ReLU \\ Conv3, BN, ReLU \end{bmatrix}$	$(24 \times 2^i) \times (160 \times 2^{-i}) \times$ $(160 \times 2^{-i}) \times (128 \times 2^{-i})$
	DownSample EncBlock	Maxpool(kernel2) $\begin{bmatrix} Conv3, BN, ReLU \\ Conv3, BN, ReLU \end{bmatrix}$	$192 \times 10 \times 10 \times 8$
Decoder: Chief	AttentionBlock i DeBlock i (i = 1, 2, 3)	AttentionLayer $\begin{bmatrix} Conv3, BN, ReLU \\ Conv3, BN, ReLU \end{bmatrix}$	$(192 \times 2^{-i}) \times (10 \times 2^i) \times$ $(10 \times 2^i) \times (8 \times 2^i)$
	AttentionBlock DeBlock	AttentionLayer $\begin{bmatrix} Conv3, BN, ReLU \\ Conv3, BN, ReLU \end{bmatrix}$	$24 \times 160 \times 160 \times 128$
	EndConv	$Conv3$	$4 \times 160 \times 160 \times 128$
Decoder: Auxiliary	Decoder	$\begin{bmatrix} GN, ReLU \\ Conv3, Dense \end{bmatrix}$	256×1
	Sample	Sample $\sim$ $\left[N(\mu(128), \sigma^2(128))\right]$	128×1
	UpBlock	$\begin{bmatrix} Dense, ReLU \\ Conv1, Uplinear \end{bmatrix}$	$256 \times 10 \times 10 \times 8$
	DeBlock i (i = 1, 2, 3)	Conv3, UpLinear $\begin{bmatrix} GN, ReLU, Conv3 \\ GN, ReLU, Conv3 \end{bmatrix}$ AddId	$(256 \times 2^{-i}) \times (10 \times 2^i) \times$ $(10 \times 2^i) \times (8 \times 2^i)$
	DeBlock	Conv3, UpLinear $\begin{bmatrix} GN, ReLU, Conv3 \\ GN, ReLU, Conv3 \end{bmatrix}$ AddId	$32 \times 160 \times 160 \times 128$
	EndConv	$Conv3$	$4 \times 160 \times 160 \times 128$

$$\mathcal{L}_{cross} = -\sum g_j^c \log(p_j^c) \tag{2}$$

The term $\mathcal{L}_{L_2}$ corresponds to an L_2 loss applied to the output $R_2 \in \mathbb{R}^{C\times H\times W\times D}$ of the VAE branch, with the objective of aligning it with the input data X. This loss mechanism operates by quantifying the Euclidean distance between the generated VAE output and the original input, fostering an optimization process that aims to minimize the dissimilarity between the two representations.

$$\mathcal{L}_{L_2} = \|R_2 - X\|_2^2 \tag{3}$$

$\mathcal{L}_{KL}$ stands for the conventional penalty term utilized in a variational autoencoder (VAE) framework. It quantifies the Kullback-Leibler (KL) divergence between the estimated normal distribution $N(\mu, \sigma^2)$ and a predefined prior distribution $N(0, 1)$. Notably, this term plays a crucial role in regulating the latent space during the VAE training process. Its closed-form expression is an essential feature of VAEs and contributes to the overall effectiveness of the model.

$$\mathcal{L}_{KL} = \frac{1}{N}\sum \mu^2 + \sigma^2 - \log\sigma^2 - 1 \tag{4}$$

The total loss function is:

$$\mathcal{L} = (1-\alpha)\times\mathcal{L}_{cross} + \mathcal{L}_{dice}\times\alpha + \alpha_1\times\mathcal{L}_{L2} + \alpha_2\times\mathcal{L}_{KL} \tag{5}$$

where $\alpha, \alpha_1, \alpha_2$ are the hyper-parameters to balance each loss item.

3 Results

3.1 Data and Implementation

Data used in this publication are provided by the FeTS Challenge, and were obtained as part of the RSNA-ASNR-MICCAI Brain Tumor Segmentation (BraTS) Challenge project through Synapse ID (syn28546456) [16, 17, 2]. The training set contains 1000 samples each with four modalities of T1, T1ce, T2 and FLAIR while the valid set contains 200 samples. Every modality of a sample is presented with a volume of 240 × 240 × 155 which is randomly cropped to 160 × 160 × 128. Although the validation set is also provided but without the ground truth. Therefore we divide the training data into two parts without overlap, and use about 1/4 data merely for model evaluation and result analysis. We use the classical Dice score (the higher the better) as the metric, calculated in regions of ET, TC and WT respectively. For the loss weights, we set $\alpha = 0.5$, $\alpha_1 = 0.5$ and $\alpha_2 = 0.5$.

Based on the open source of [23], the network is implemented under the Pytorch framework. We train it with one NVIDIA A100 GPUs (each has 80GB memory) from scratch using a batch size of 1. The initial learning rate is $4\times e^{-4}$. For more training details including learning rate decay and data augmentation strategies, please refer to [23]. The network details are provided in Table 1. $Conv3$

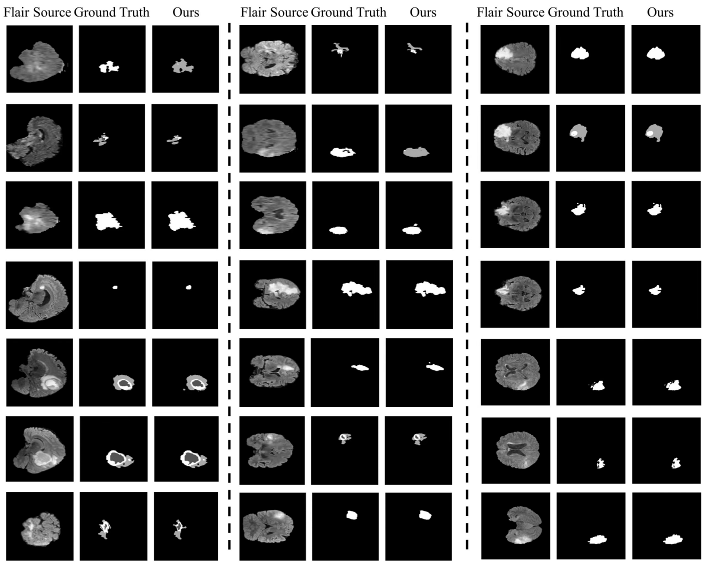


Fig. 2. The segmentation results of our method.

denotes a convolutional layer with the kernel size of 3 × 3 × 3, BN is short for Batch Normalization, GN is short for Group Normalization, $UpLinear$ means 3D linear spatial upsampling, $Dense$ stands for full connections and $AddId$ represents addition of identity skip connection.

3.2 Discussion

Following the acquisition of the segmentation outcomes, we proceed with an evaluation encompassing per-class Dice coefficients, Hausdorff distances.

Dice coefficient, often denoted as D, is a common metric used for assessing the overlap between two sets. It is defined as:

$$\mathcal{D} = \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{6}$$

Where X represents the ground truth segmentation mask and Y represents the predicted segmentation mask. $|\cdot|$denotes the cardinality of a set, and$|\cdot \cap \cdot|$ represents the intersection of two sets.

Hausdorff distance, denoted as H, is a measure of the maximum distance between the points of two sets. Specifically, the Hausdorff distance between sets X and Y is defined as:

$$H(X, Y) = \{sup_{x \in X}\ inf_{y \in Y} d(x,y), sup_{y \in Y}\ inf_{x \in X} d(x,y)\} \tag{7}$$

Where $d(x,y)$ represents the distance between point x in set X and point y in set Y. The Hausdorff distance measures the similarity between two sets by capturing the maximum distance of a point in one set to the closest point in the other set.

Besides, two innovative performance metrics referred to as lesion-wise Dice scores and lesion-wise Hausdorff distances at the 95th percentile (HD95). These were developed to evaluate segmentation performance at a lesion level rather than at the whole study level. By evaluating segmentation performance at the lesion level we can understand how well models detect and segment multiple individual lesions within a single patient. Traditional performance metrics used in prior BraTS are biased for large lesions. The results of this evaluation are presented in both Table 2 and Table 3. A comparison against models with a single decoder branch highlights a noticeable enhancement in segmentation performance with the integration of the VAE. This enhancement serves as a compelling testament to the efficacy of our proposed methodology. For enhanced clarity, we visually represent the segmentation outcomes in Figure 2. In contrast to the quantitative metrics, these visual depictions offer a more intuitive assessment of segmentation quality. Across most scenarios, our outcomes closely resemble the ground truth, demonstrating the promising practical applicability of our approach. However, it is acknowledged that certain shortcomings persist within the segmentation results, such as the occasional omission of small, scattered regions. We proceed with a comprehensive examination of the convergence within our network. We present the curves depicting the variation in loss across training iterations in Figure 3, providing a visual representation of the model's convergence dynamics. Notably, a rapid reduction in loss is evident during the initial 50 iterations.

Table 2. Dice score and Hausdorff distance-95 (HD95) measurements of the proposed segmentation method. EN - enhancing tumor core, WT - whole tumor, TC - tumor core.

Method	Dice Score (%) ↑			HD95 Score ↓		
	ET	WT	TC	ET	WT	TC
Attention+VAE	75.5	80.2	70.8	26.49	13.17	31.73
Attention	72.9	78.7	69.4	31.31	15.14	40.22

Table 3. Lesion-wise dice score and lesion-wise Hausdorff distance-95 measurements of the proposed segmentation method.

Method	LesionWise Dice (%) ↑			LesionWise HD95 ↓		
	ET	WT	TC	ET	WT	TC
Attention+VAE	62.6	72.4	62.8	95.11	59.39	85.37
Attention	59.7	71.9	56.5	109.78	63.92	106.26

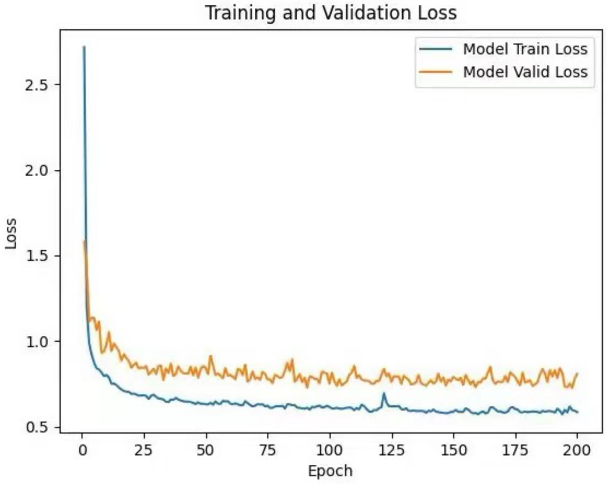


Fig. 3. The loss decline trend of the model in train and valid datasets.

4 Conclusion

In this paper, we have taken steps to elevate the capabilities of the encoder within the encoder-decoder network architecture by incorporating 3D convolutions and attention gates. This augmentation has resulted in remarkable brain tumor segmentation outcomes when applied to MRI samples. This enhancement offers a threefold advantage: (1) Exploiting 3D Convolution: The integration of 3D convolution goes beyond capturing local correlations within individual modalities of MRI. It extends its reach across all four modalities, comprehensively enhancing our ability to decipher intricate patterns. (2) Harnessing Attention Gates: Our implementation of attention gates has proven to be a pivotal advancement. These gates facilitate superior feature fusion through skip connections, effectively thwarting the risk of detail leakage. (3) Empowering Encoder with VAE: We have also integrated a Variational Autoencoder (VAE) as an auxiliary decoder, effectively bolstering the capabilities of the encoder to capture complex features. The entire network has been meticulously trained and validated using MRI data provided by the esteemed Federated Tumor Segmentation (FeTS) 2023 Challenge. Notably, our approach excels in producing convincing and promising segmentation outcomes in the Federated Evaluation phase, firmly underscoring the remarkable generalization prowess of our proposed methodology.

5 Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (No. 62106037, No. 62076052), in part by the Major Program of the National Social Science Foundation of China (No.19ZDA127), and in part by the Fundamental Research Funds for the Central Universities (No. DUT22YG205).

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: IEEE International Conference on Computer Vision (ICCV). pp. 6836–6846 (2021)
2. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
3. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C.: Segmentation labels and radiomic features for the preoperative scans of the tcga-gbm collection (2017). DOI: https://doi. org/10.7937 K **9**
4. 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)
5. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C.: Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection. The cancer imaging archive **286** (2017)
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision (ECCV). pp. 213–229. Springer (2020)
7. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: International Conference on Machine Learning. pp. 1691–1703. PMLR (2020)
8. 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)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Karargyris, A., Umeton, R., Sheller, M.J., Aristizabal, A., George, J., Wuest, A., Pati, S., Kassem, H., Zenk, M., Baid, U., et al.: Federated benchmarking of medical artificial intelligence with medperf. Nature Machine Intelligence pp. 1–12 (2023)
11. Li, Y., Xie, X., Fu, H., Luo, X., Guo, Y.: A compact transformer for adaptive style transfer. In: 2023 IEEE International Conference on Multimedia and Expo (ICME). pp. 2687–2692. IEEE (2023)
12. 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)
13. 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)
14. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)

15. Ostrom, Q.T., Gittleman, H., Fulop, J., Liu, M., Blanda, R., Kromer, C., Wolinsky, Y., Kruchko, C., Barnholtz-Sloan, J.S.: Cbtrus statistical report: primary brain and central nervous system tumors diagnosed in the united states in 2008-2012. Neuro-oncology **17**(suppl_4), iv1–iv62 (2015)
16. Pati, S., Baid, U., Zenk, M., Edwards, B., Sheller, M., Reina, G.A., Foley, P., Gruzdev, A., Martin, J., Albarqouni, S., et al.: The federated tumor segmentation (fets) challenge. arXiv preprint arXiv:2105.05874 (2021)
17. Reina, G.A., Gruzdev, A., Foley, P., Perepelkina, O., Sharma, M., Davidyuk, I., Trushkin, I., Radionov, M., Mokrov, A., Agapov, D., et al.: Openfl: An open-source framework for federated learning. arXiv preprint arXiv:2105.06413 (2021)
18. 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)
19. Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D.: Attention gated networks: Learning to leverage salient regions in medical images. Medical image analysis **53**, 197–207 (2019)
20. Su, R., Yu, Q., Xu, D.: Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In: IEEE International Conference on Computer Vision (ICCV). pp. 1533–1542 (2021)
21. Tang, Z., Liao, Y., Liu, S., Li, G., Jin, X., Jiang, H., Yu, Q., Xu, D.: Human-centric spatio-temporal video grounding with visual transformers. IEEE Transactions on Circuits and Systems for Video Technology (2021)
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances Neural Information Processing Systems (NeurIPS) **30** (2017)
23. 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)
24. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
25. Xie, X., Li, Y., Huang, H., Fu, H., Wang, W., Guo, Y.: Artistic style discovery with independent components. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19870–19879 (June 2022)
26. Zhao, L., Cai, D., Sheng, L., Xu, D.: 3dvg-transformer: Relation modeling for visual grounding on point clouds. In: IEEE International Conference on Computer Vision (ICCV). pp. 2928–2937 (2021)
27. Zhao, L., Guo, J., Xu, D., Sheng, L.: Transformer3d-det: Improving 3d object detection by vote refinement. IEEE Transactions on Circuits and Systems for Video Technology **31**(12), 4735–4746 (2021)
28. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 3–11. Springer (2018)