A Tri-Attention fusion guided multi-modal segmentation network

Tongxue Zhou, Su Ruan, Pierre Vera, Stéphane Canu

Version of Record: https://www.sciencedirect.com/science/article/pii/S0031320321005938
Manuscript_2f284cdf6236a50943974b444f587c3c

A Tri-attention Fusion Guided Multi-modal Segmentation Network

Tongxue Zhou[a,b,c], Su Ruan[a,c,*], Pierre Vera[d], Stéphane Canu[b,c]

[a] *Université de Rouen Normandie, LITIS - QuantIF, Rouen 76183, France*
[b] *INSA de Rouen, LITIS -Apprentissage, Rouen 76800, France*
[c] *Normandie Univ, INSA Rouen, UNIROUEN, UNIHAVRE, LITIS, France*
[d] *Department of Nuclear Medicine, Henri Becquerel Cancer Center, Rouen, 76038, France*

Abstract

In the field of multimodal segmentation, the correlation between different modalities can be considered for improving the segmentation results. Considering the correlation between different MR modalities, in this paper, we propose a multi-modality segmentation network guided by a novel tri-attention fusion. Our network includes N model-independent encoding paths with N image sources, a tri-attention fusion block, a dual-attention fusion block, and a decoding path. The model independent encoding paths can capture modality-specific features from the N modalities. Considering that not all the features extracted from the encoders are useful for segmentation, we propose to use dual attention based fusion to re-weight the features along the modality and space paths, which can suppress less informative features and emphasize the useful ones for each modality at different positions. Since there exists a strong correlation between different modalities, based on the dual attention fusion block, we propose a correlation attention module to form the tri-attention fusion block. In the correlation attention module, a correlation description block is first used to learn the correlation between modalities and then a constraint based on the correlation is used to guide the network to learn the latent correlated features which are more relevant for segmentation. Finally, the obtained fused feature repre-

*Corresponding author
Email address: su.ruan@univ-rouen.fr (Su Ruan)

sentation is projected by the decoder to obtain the segmentation results. Our experiment results tested on BraTS 2018 dataset for brain tumor segmentation demonstrate the effectiveness of our proposed method.

Keywords: Multi-modality fusion, Correlation, Brain tumor segmentation, Deep learning

1. Introduction

Multimodal segmentation using a single model remains challenging due to the different image characteristics of different modalities. A key challenge is to exploit the latent correlation between modalities and to fuse the complementary information to improve the segmentation performance. In this paper, we proposed a method to exploit the multi-source correlation and apply it to brain tumor segmentation task.

Magnetic Resonance Imaging (MRI) is commonly used in radiology to diagnose brain tumors, it is a non-invasive and good soft tissue contrast imaging modality, which provides invaluable information about shape, size, and localization of brain tumors without exposing the patient to a high ionization radiation [1, 2, 3]. The commonly used sequences are T1-weighted (T1), contrast-enhanced T1-weighted (T1c), T2-weighted (T2) and Fluid Attenuation Inversion Recovery (FLAIR) images. In this work, we refer to these images of different sequences as modalities. Different modalities can provide complementary information to analyze different subregions of gliomas. For example, T2 and FLAIR highlight the tumor with peritumoral edema, designated whole tumor. T1 and T1c highlight the tumor without peritumoral edema, designated tumor core. An enhancing region of the tumor core with hyper-intensity can also be observed in T1c, designated enhancing tumor core. Therefore applying multi-modal images can reduce the information uncertainty and improve clinical diagnosis and segmentation accuracy.

Inspired by a fact that, there is strong correlation between multi MR modalities, since the same scene (the same patient) is observed by different modalities

[4, 5]. We propose a novel tri-attention fusion to guide 3D multi-modal brain tumor segmentation network. A preliminary conference version appeared at MICCAI 2020 [5], which focused on the multi-modal brain tumor segmentation with missing modality. This journal version extended previous work, and applied it on the full modality brain tumor segmentation. The main contributions in this paper are: 1) A novel correlation description block is introduced to discover the latent multi-source correlation between modalities. 2) A correlation constraint using KL divergence is proposed to aide the segmentation network to extract the correlated feature representation for a better segmentation. 3) A tri-attention fusion strategy is proposed to re-weight the feature representation along modality-attention, spatial-attention and correlation-attention paths. 4) The first 3D multimodal brain tumor segmentation network guided by tri-attention fusion is proposed.

The rest of the paper is organised as follows. Section 2 reviews the relevant prior work, Section 3 details our proposed method, Section 4 describes the data used and implementation details, Section 5 presents the experiment results, Section 6 gives a further discussion about our method, and Section 7 concludes our work.

2. Related work

A number of conventional brain tumor segmentation approaches haven been presented in recent years, including probability theory [4], kernel feature selection [6], belief function [7], random forests [8], conditional random fields [9] and support vector machines [10]. However, the performance is limited due to the complex brain anatomy structure, different shape, texture of gliomas, and the low contrast of MR images (see Figure 1).

Recently, various deep learning-based approaches have been successfully designed for brain tumor segmentation. Wang et al. [11] developed a cascaded system to segment brain tumor using three binary segmentation subnetworks. Chen et al. [12] proposed a novel deep convolutional symmetric neural network,

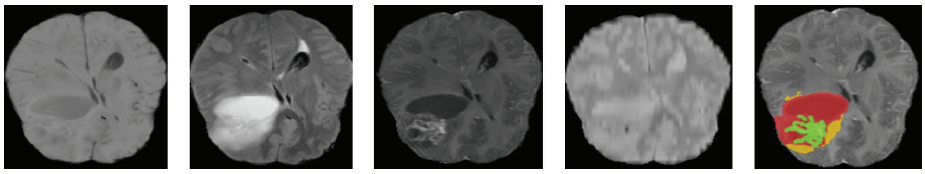

Figure 1: Example of data from a training subject. The first four images from left to right show the MRI modalities: T1-weighted (T1), Fluid Attenuation Inversion Recovery (FLAIR), contrast enhanced T1-weighted (T1c), T2-weighted (T2) images, and the fifth image is the ground truth labels created by experts. The color is used to distinguish the different tumor regions: red: necrotic and non-enhancing tumor, yellow: edema, green: enhancing tumor, black: healthy tissue and background.

which combines the symmetry prior knowledge into brain tumor segmentation. Zhao et al. [13] proposed a deep learning model integrating FCNNs and CRFs for brain tumor segmentation. Myronenko et al. [14] proposed a segmentation network for brain tumor from multimodal 3D MRIs, where variational auto-encoder branch is added into the U-net to further regularize the decoder in the presence of limited training data. Wei et al. [15] proposed a multi-model, multi-size and multi-view deep model for brain tumor segmentation. Dolz et al. [16] presented an ensemble of deep CNNs to segment isointense infant brains in multi-modal MRI images. Chen et al. [17] proposed a dual-force training strategy to explicitly encourage deep models to learn high-quality multi-level features for brain tumor segmentation.

For multi-modal segmentation task, exploiting the complimentary information from different modalities plays an essential role in the final segmentation accuracy. The single-encoder-based method and multi-encoder-based method are the common used network frameworks [18]. The single-encoder-based method [11, 19] directly fuses the different multi-source images in the input space, while the correlations between different modalities are not well exploited. However, the multi-encoder-based method [20], applied separate encoders to extract individual feature representations, respectively. And it can achieve a better segmentation result than the former one [21]. Since the effective feature representation can attribute to a better segmentation performance. Inspired by the atten-

tion mechanism [22], in this paper, we first proposed a dual attention based fusion block to selectively emphasize feature representations, which consists of a modality attention module and a spatial attention module. The proposed fusion block uses the individual features obtained from encoders to derive a modality-wise and a spatial-wise weight map that quantify the relative importance of each modality's features and also of the different spatial locations in each modality. These fusion maps are then multiplied with the individual feature representations to obtain a fused feature representation of the complementary multi-modality information. In this way, we can discover the most relevant characteristics to aide the segmentation.

For multi-modal MR brain tumor segmentation, since the four MR modalities are from the same patient, there exists a strong correlation in the tumor regions between modalities [4]. Therefore, we proposed a correlation attention module, it consists of a correlation description block and a KL divergence based correlation constraint. It can exploit and utilize the correlation between modalities to improve the segmentation performance. In the correlation attention module, a correlation description block is first used to exploit the correlation between the spatial-attention feature representations, and then a correlation constraint based on KL divergence is used to guide the segmentation network to learn the correlated features to enhance the segmentation result. The novelty of this method is capable of exploiting and utilizing the latent multi-source correlation to help the segmentation. The proposed method can be generalized to other applications.

3. Multi-modal segmentation with correlation costraints

In this paper, we aim to exploit the multi-source correlation between modalities and utilize the correlation to constrain the network to learn more effective feature so as to improve the segmentation performance. To learn complementary features and cross-modal inter-dependencies from multi-modality MRIs, we applied the multi-encoder based framework. It takes 3D MRI modality as

input in each encoder. Each encoder can produce a modality-specific feature representation. At the lowest level of the network, the tri-attention fusion block is used, which includes a dual-attention fusion block and a correlation attention module. The dual-attention fusion block can re-weight the feature representation along modality-wise and spatial-wise. The correlation attention module is to first exploit the latent multi-source correlation between the spatial-attention feature representations. Then, it uses a correlation based constraint to guide the network to learn the effective feature information. Finally, the fused feature representation is projected by decoder to the label space to obtain the segmentation result. The overview of the proposed network is described in Figure 2.

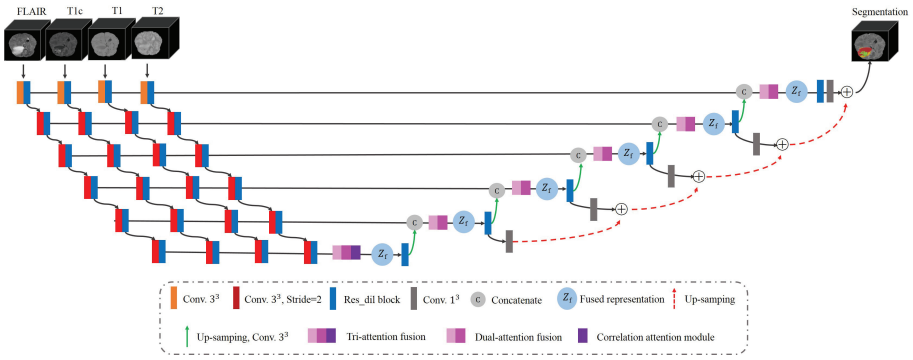

Figure 2: Overview of our proposed segmentation network. The backbone is a multi-encoder based 3D U-Net, the separate encoders enable the network to extract the independent feature representations. The proposed dual-attention fusion block is to re-weight the feature representations along modality and space paths. The tri-attention fusion block consists of the dual-attention fusion and a correlation attention module.

3.1. Encoder and Decoder

It's likely to require different receptive fields when segmenting different regions in an image, a standard U-Net can't get enough semantic features due to the limited receptive field. Inspired by dilated convolution, we use residual block with dilated convolutions (rate = 2, 4) (res_dil block) on both encoder part and decoder part to obtain features at multiple scale. The encoder includes a

convolutional block, a res_dil block followed by skip connection. All convolutions are $3 \times 3 \times 3$. Each decoder level begins with up-sampling layer followed by a convolution to reduce the number of features by a factor of 2. Then the upsampled features are combined with the features from the corresponding level of the encoder part using concatenation. After the concatenation, we use the res_dil block to increase the receptive field. In addition, we employ deep supervision [23] for the segmentation decoder by integrating segmentation results from different levels to form the final network output.

3.2. Tri-attention Fusion Strategy

The purpose of fusion is to stand out the most important features from different source images to highlight regions that are greatly relevant to the target region. Since different MR modalities can identify different attributes of the target tumor. In addition, from the same MR modality, we can learn different content at different positions. Inspired by the attention mechanism [22], we propose a dual-attention fusion block to enable a better integration of the complementary information between modalities, which consists of a modality attention module, and a spatial attention module.

Inspired by a fact that, there is strong correlation between multi MR modalities, since the same brain tumor region is observed by different modalities [4]. From Figure 3 presenting joint intensities of the MR images, we can observe a strong correlation (not always linear) in intensity distribution between each pair of modalities. To this end, it's reasonable to assume that a strong correlation also exists in latent feature representation between modalities. Therefore, we proposed a correlation attention module and integrated it to the dual-attention fusion block to achieve a tri-attention fusion block. It's used to exploit and utilize the multi-source correlation between modalities, the architeccture is depicated in Figure 4.

The input modality $\{X_i, ..., X_n\}$, where $n = 4$, is first input to the independent encoders (with learning parameters θ including the number of the filters and dropout rate) to learn the modality-specific representation Z_i. Then, a dual-

attention fusion block is used. It takes the concatenation of the independent feature representations as input to produce the modality-weight and spatial-weight, respectively. And the two weights are multiplied with the input feature representation to obtain the modality-attention feature representation Z_{im} and spatial-attention feature representation Z_{is}, respectively. Finally, the learned fused feature representation is obtained by adding the modality-attention feature representation and spatial-attention feature representation.

The obtained spatial-attention feature representation Z_{is} is passed to the Correlation Description (CD) block consisting of two fully connected layers and LeakyReLU, it maps the spatial-attention feature representation Z_{is} to a set of independent parameters $\Gamma_i = \{\alpha_i, \beta_i, \gamma_i\},\ i = 1, ..., n$. Finally, the correlated representation of i modality F_i can be obtained via correlation expression (Equation 1).

$$F_i = \alpha_i \odot Z_{is}^2 + \beta_i \odot Z_{is} + \gamma_i \tag{1}$$

It is noted that the nonlinear correlation expression we proposed in this work is specific to our work. However, the proposed correlation description block can be generally integrated to any multi-source correlation problem, and the specific correlation expression will depend on the application. In addition, we compare and discuss why the simplest linear correlation expression is not good for this work in Section 6.1.

Then, the Kullback–Leibler divergence (Equation 2) is used to measure the divergence between the estimated correlated feature representation of i modality and the spatial-attention feature representation of j modality, which enables the segmentation network to learn the latent correlated features which are more relevant for segmentation. To make it clear, we take T1 modality (X_1) and T1c modality (X_3) as example, since there exists a correlation between the two modalities, the spatial attention module is first used to obtain the two spatial-attention feature representations of T1 modality (Z_{1s}) and T1c modality (Z_{3s}), then the correlated feature representation (F_1) of modality T1 can be obtained

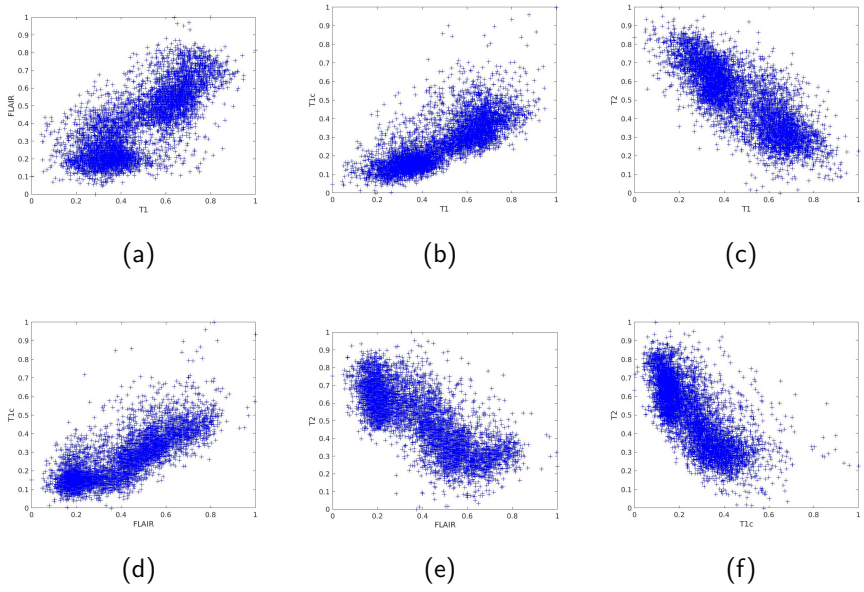

Figure 3: Joint intensity distributions of MR images: (a) T1-FLAIR, (b) T1-T1c, (c) T1-T2, (d) FLAIR-T1c, (e) FLAIR-T2, (f) T1c-T2. The intensity of the first modality is read on abscissa axis and that of the second modality on the ordinate axis.

by CD block and Equation 1. Finally, the KL based correlation loss function is applied to constrain the two distributions (F_1 and Z_{3s}) to be as close as possible.

$$L_{correlation} = \sum_{x \in X} P(x) log \frac{P(x)}{Q(x)} \tag{2}$$

where P and Q are probability distributions of spatial-attention feature representation (Z_{js}) of modality j and correlated feature representation(F_i) of modality i, ($i \neq j$), respectively, which defined on the same probability space X.

From Figure 4, we can observe the characteristics of the target tumors in the four independent feature representations (Z_1, Z_2, Z_3, Z_4) are not obvious. However, the modality attention module can stand out the different attributes of the modalities to provide complementary information. For example, the FLAIR modality (Z_{2m}) highlights the whole tumor region and T1c modality (Z_{3m}) stands out the tumor core region (red and green). In the spatial attention

module, all the positions related to the target tumor regions are highlighted. In this way, we can discover the most relevant characteristics between modalities. Furthermore, the proposed tri-attention fusion strategy can be directly adapted to any multi modal (if existing a correlation relationship) fusion problem.

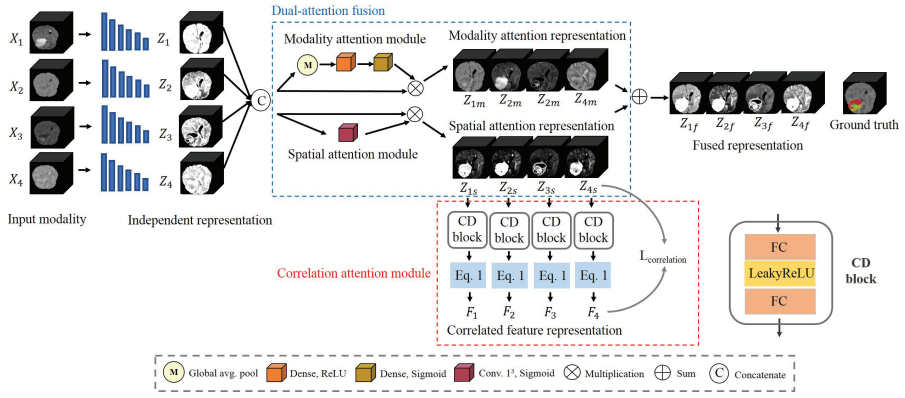

Figure 4: Architecture of the tri-attention fusion strategy. The individual feature representations (Z_1, Z_2, Z_3, Z_4) are first concatenated, then they are re-weighted by dual-attention fusion block along modality attention module and spatial attention module to achieve the modality attention representation Z_{im} and spatial attention representation Z_{is}. In addition, the correlation attention module is used to constrain the spatial-attention representations to learn segmentation-related representation. Finally, the Z_{im} and Z_{is} are added to obtain the fused feature representation Z_{if}.

4. Data and Implementation Details

4.1. Data

The datasets used in the experiments come from BraTS 2018 dataset. The training set includes 285 patients, each patient has four image modalities including T1, T1c, T2 and FLAIR. Following the challenge, four intra-tumor structures have been grouped into three mutually inclusive tumor regions: (a) whole tumor (WT) that consists of all tumor tissues, (b) tumor core (TC) that consists of the enhancing tumor, necrotic and non-enhancing tumor core, and (c) enhancing tumor (ET). The provided data have been pre-processed by organisers: co-registered to the same anatomical template, interpolated to the

same resolution ($1mm^3$) and skull-stripped. The ground truth have been manually labeled by experts. We did additional pre-processing with a standard procedure. The N4ITK [24] method is used to correct the distortion of MRI data, and intensity normalization is applied to normalize each modality of each patient. To exploit the spatial contextual information of the image, we use 3D images, crop and resize them from $155 \times 240 \times 240$ to $128 \times 128 \times 128$.

4.2. Implementation Details

Our network is implemented in Keras with a single Nvidia GPU Quadro P5000 (16G). The models are optimized using the Adam optimizer(initial learning rate = 5e-4) with a decreasing learning rate factor 0.5 with patience of 10 epochs, to avoid over-fitting, early stopping is used when the validation loss isn't improved for 50 epoch. We randomly split the dataset into 80% training and 20% testing.

4.3. The choices of loss function

For segmentation, we use dice loss to evaluate the overlap rate of prediction results and ground truth.

$$L_{dice} = 1 - 2\frac{\sum_{i=1}^{C}\sum_{j=1}^{N} p_{ij} g_{ij} + \epsilon}{\sum_{i=1}^{C}\sum_{j=1}^{N} (p_{ij} + g_{ij}) + \epsilon} \tag{3}$$

where N is the set of all examples, C is the set of the classes, p_{ij} is the probability that pixel i is of the tumor class j, the same is true for g_{ij}, and ϵ is a small constant to avoid dividing by 0.

The network is trained by the overall loss function as follow:

$$L_{total} = L_{dice} + \lambda \sum_{i=1}^{n} L_{correlation_n} \tag{4}$$

where λ is the trade-off parameter weighting the importance of each component, n denotes the number of correlation pair, in this work, we used three most correlated pairs: T1-T1c, T1-T2, T2-FLAIR.

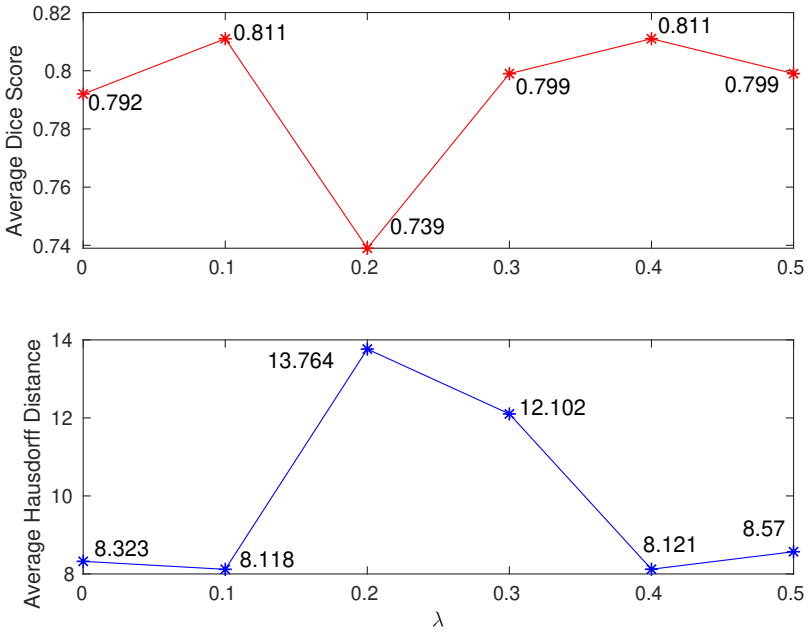

Figure 5: Comparison of different weight coefficients in the loss function. Average Dice Score vs λ and Average Hausdorff Distance vs λ.

We did a grid search between $(0, 0.5)$ to determine the optimal value for the weight coefficient λ, Figure 5 shows the comparison of average Dice Score and Hausdorff Distance between different weight coefficients, we found that $\lambda = 0.1$ can achieve the best segmentation results.

4.4. Evaluation metrics

To evaluate the proposed method, two evaluation metrics: Dice Score and Hausdorff distance are used to obtain quantitative measurements of the segmentation accuracy:

1) Dice Score: It is designed to evaluate the overlap rate of prediction results and ground truth. It ranges from 0 to 1, and the better predict result will have a larger Dice value.

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{5}$$

where TP represents the number of true positive voxels, FP represents the number of false positive voxels, and FN represents the number of false negative voxels.

2) Hausdorff Distance (HD): It is computed between boundaries of the prediction results and ground-truth, it is an indicator of the largest segmentation error. The better predict result will have a smaller HD value.

$$HD = \max\{sup_{r \in \partial R} d_m(s, r), sup_{s \in \partial S} d_m(r, s)\} \tag{6}$$

where ∂S and ∂R are the sets of tumor border voxels for the predicted and the real annotations, and $d_m(v, v)$ is the minimum of the Euclidean distances between a voxel v and voxels in a set v.

5. Experiment Results

We conduct a series of comparative experiments to demonstrate the effectiveness of our proposed method and compare it to other approaches. In Section 5.1.1, we first perform an ablation experiment to see the importance of our proposed components and demonstrate that adding the proposed components can enhance the segmentation performance. In Section 5.1.2, we compare our method with the state-of-the-art U-Net based methods. In Section 5.2, the qualitative experiment results further demonstrate that our proposed method can achieve a promising segmentation result.

5.1. Quantitative analysis

To prove the effectiveness of our network, we first did an ablation experiment to see the effectiveness of our proposed components, and then we compare our method with the state-of-the-art methods. All the results are obtained by online evaluation platform[1].

[1] https://ipp.cbica.upenn.edu/

5.1.1. Effectiveness of individual modules

To assess the performance of our method, and see the importance of the proposed components in our network, including dual attention fusion strategy and correlation attention module, we did an ablation experiment, our network without the dual attention fusion and correlation attention module is denoted as baseline. From Table 1, we can observe the baseline method achieves Dice Score of 0.726, 0.867, 0.764 for enhancing tumor, whole tumor, tumor core, respectively. When the dual attention fusion strategy is applied to the network, we can see an increase of Dice Score and Hausdorff Distance across all tumor regions with an average improvement of 0.76% and 6.44% compared to the baseline, respectively. The major reason is that the proposed fusion block can help to emphasize the most important representations from the different modalities across different positions in order to boost the segmentation result. In addition, another advantage of our method is using the correlation attention module in the lowest layer, which can constrain the encoders to discover the latent multi-source correlation representation between modalities and then guide the network to learn correlated representation to achieve a better segmentation. From the results, we can observe that with the assistance of correlation attention module, the network can achieve the best Dice Score of 0.75, 0.887 and 0.796 and Hausdorff Distance of 7.687, 8.306 and 8.362 for enhancing tumor, whole tumor, tumor core, respectively with an average improvement of 3.18% and 8.75% relating to the baseline. The results in Table 1 demonstrate the effectiveness of each proposed component and our proposed network architecture can perform well on brain tumor segmentation.

5.1.2. Comparisons with the state-of-the-art

Since access to the testing set of BraTS 2018 was closed after the challenge, we compare our proposed method with the state-of-the-art methods on BraTS 2018 online validation set, which contains 66 images of patients without the ground-truth. We first predict the segmentation results on local machine and then submitted on the online evaluation platform to obtain the evaluation

Table 1: Evaluation of our proposed method on Brats 2018 local test dataset, (1) Baseline (2) Baseline + Dual attention fusion (3) Baseline + Tri-attention fusion, ET, WT, TC denote enhancing tumor, whole tumor and tumor core, respectively, bold results denotes the best scores.

Methods	Dice Score				Hausdorff (mm)			
	ET	WT	TC	Avg	ET	WT	TC	Avg
(1)	0.726	0.867	0.764	0.786	8.743	8.463	9.482	8.896
(2)	0.733	0.879	0.765	0.792	8.003	**7.813**	9.153	8.323
(3)	**0.750**	**0.887**	**0.796**	**0.811**	**7.687**	8.306	**8.362**	**8.118**

results. Table 2 shows the comparison results.

(1) Zhao et al. [13] proposed to integrate Fully Convolutional Neural Networks (FCNNs) and Conditional Random Fields (CRFs) in a unified framework, where three segmentation models using 2D image patches and slices are trained in axial, coronal and sagittal views respectively, and they are combined to segment brain tumors using a voting based fusion strategy.

(2) Kamnitsas et al. [19] introduced EMMA, an ensemble of widely varying CNNs. By combining a heterogeneous collection of networks, the proposed model is insensitive to independent failures of CNN components and thus generalises well, which won the first place in BraTS 2017 challenge.

(3) Hu et al. [25] proposed the multi-level up-sampling network (MU-Net) for automated segmentation of brain tumors, where a novel global attention (GA) module is used to combine the low level feature maps obtained by the encoder and high level feature maps obtained by the decoder.

(4) Gates et al. [26] applied a multi-scale convolutional neural network based on the DeepMedic [27] to segment brain tumor.

(5) Tuan et al. [28] proposed using Bit-plane to generate a series of binary images by determining significant bits. Then, the first U-Net used the significant bits to segment the tumor boundary, and the other U-Net utilized the original

images and images with least significant bits to predict the label of all pixel inside the boundary.

(6) Hu et al. [29] introduced the 3D-residual-Unet architecture. The network comprises a context aggregation pathway and a localization pathway, which encoder abstract representation of the input, and then recombines these representations with shallower features to precisely localize the interest domain via a localization path.

(7) Myronenko et al. [14] proposed a 3D MRI brain tumor segmentation using autoencoder regularization, where a variational autoencoder branch is added to reconstruct the input image itself in order to regularize the shared decoder and impose additional constraints on its layers.

The best result in BraTS 2018 Challenge is from [14], which achieves 0.814, 0.904 and 0.859 in terms of Dice Score on enhancing tumor, whole tumor and tumor core regions, respectively. However, it uses 32 initial convolution filters and a lot of memories (NVIDIA Tesla V100 32GB GPU is required) to train the model, which is computationally expensive. While our method used only 8 initial filters, and from Table 2, it can be observed that our proposed method can yield a competitive results in terms of Dice Score and Hausdorff Distance across all the tumor regions. Compared with other methods, [29] has a better Dice Score on enhancing tumor, while our method achieves a better average Dice Score on all the tumor regions with an improvement of 3.84%, and it can also obtain an average improvement of 7.5% for Hausdorff Distance.

5.2. Qualitative analysis

In order to evaluate the robustness of our model, we randomly select several examples on BraTS 2018 dataset and visualize the segmentation results in Figure 6. From Figure 6, we can observe that the segmentation results are gradually improved when the proposed strategies are integrated, these comparisons indicate that the effectiveness of the proposed strategies. In addition, with all the proposed strategies, our proposed method can achieve the best results.

Table 2: Comparison of different methods on BraTS 2018 validation dataset, ET, WT, TC denote enhancing tumor, whole tumor, tumor core, respectively, bold results denotes the best scores, underline results denotes the second best results.

Methods	Dice Score				Hausdorff (mm)			
	ET	WT	TC	Avg	ET	WT	TC	Average
[13]	0.62	0.84	0.73	0.715	-	-	-	-
[19]	0.629	0.847	0.67	0.715	-	-	-	-
[25]	0.69	0.88	0.74	0.77	6.69	4.76	10.67	<u>7.373</u>
[26]	0.678	0.805	0.685	0.723	14.522	14.415	20.017	16.318
[28]	0.682	0.818	0.699	0.733	7.016	9.421	12.462	9.633
[29]	<u>0.719</u>	0.856	0.769	0.781	<u>5.5</u>	10.843	<u>9.985</u>	8.776
[14]	**0.814**	**0.904**	**0.859**	**0.859**	**3.804**	**4.483**	**8.278**	**5.521**
Proposed	0.688	<u>0.876</u>	<u>0.784</u>	<u>0.783</u>	6.900	<u>6.551</u>	10.199	7.883

6. Discussion

We discuss our method from the following aspects to further demonstrate the effectiveness of our method. Initially, we explore and compare the different correlation expressions in the correlation description block to determine which functional form provides the best fit in Section 6.1. Subsequently, we analyzed the performance of correlation attention module in different layer of network in Section 6.2. Finally, we visualize the feature maps of different approaches in Section 6.3 to demonstrate the proposed fusion strategy can improve the segmentation.

6.1. Performance analysis on correlation expression

Table 3 compares the performance between linear (Equation 7) and non-linear (Equation 1) correlation expression for segmenting brain tumors. As we can see, the nonlinear correlation expression exhibits clear advantages over the

Table 3: Comparison of segmentation accuracy of different correlation expression, bold results denotes the best scores.

Methods	Dice Score				Hausdorff (mm)			
	ET	WT	TC	Avg	ET	WT	TC	Avg
Linear	0.736	0.883	0.767	0.795	8.827	8.485	9.354	8.889
Nonlinear	**0.750**	**0.887**	**0.796**	**0.811**	**7.687**	**8.306**	**8.362**	**8.118**

linear correlation expression across all the tumor regions. We explained that the capability is attributed to the complex nonlinear expression, which uses more parameters to fit a feature distribution, giving a better description for the feature distributions so as to guide the network to learn more correlated feature representations for segmentation.

$$F_i = \alpha_i \odot Z_{is} + \gamma_i \tag{7}$$

6.2. Performance analysis on correlation attention module

While experimenting with the network architectures, we have tested the addition of the correlation attention module in different layer of network. Table 4 shows the comparison results, (0) is our method without correlation attention module, which is used as a comparison baseline. As we can see, while setting the correlation attention module in the 4th and 6th layer can achieve a better segmentation results. Then we experimented to set the correlation attention module in both 4th and 6th layers ((7)), while the results aren't improved, therefore, we choose to put it in the 6th layer. Then we tried to put the correlation attention module in more layers, while the correlation attention module in multi-shallower layers ((8)-(12)) did not further improve the segmentation performance. We explained that since each layer represents different abstract feature representation of the input, where deeper levels provide more complex and abstract features, the correlation attention module can guide the most ab-

Table 4: Comparison of segmentation accuracy of correlation attention module in different layer of the network.

	Methods						Dice Score				Hausdorff (mm)			
	1st	2nd	3rd	4th	5th	6th	ET	WT	TC	Avg	ET	WT	TC	Avg
(0)	–	–	–	–	–	–	0.733	0.879	0.765	0.792	8.003	7.813	9.153	8.323
(1)	√	×	×	×	×	×	0.733	0.868	0.744	0.782	8.65	7.603	9.641	8.631
(2)	×	√	×	×	×	×	0.74	0.878	0.761	0.793	7.978	8.168	9.404	8.517
(3)	×	×	√	×	×	×	0.741	0.877	0.772	0.797	6.43	**6.994**	8.119	7.181
(4)	×	×	×	√	×	×	**0.762**	0.886	0.776	0.808	**5.906**	7.516	**7.809**	**7.077**
(5)	×	×	×	×	√	×	0.739	**0.889**	0.767	0.798	8.071	8.266	10.181	8.839
(6)	×	×	×	×	×	√	0.75	0.887	**0.796**	**0.811**	7.687	8.306	8.362	8.118
(7)	×	×	×	√	×	√	0.754	0.886	0.778	0.806	7.677	8.206	9.18	8.354
(8)	×	×	×	×	√	√	0.754	0.887	0.785	0.809	7.674	7.643	8.696	8.004
(9)	×	×	×	√	√	√	0.682	0.843	0.725	0.75	10.282	10.161	11.271	10.571
(10)	×	×	√	√	√	√	0.695	0.822	0.699	0.739	10.713	15.685	12.189	12.863
(11)	×	√	√	√	√	√	0.702	0.848	0.713	0.754	9.516	9.449	10.64	9.868
(12)	√	√	√	√	√	√	0.536	0.724	0.406	0.555	17.102	30.667	21.359	23.043

stract feature distribution to satisfy the correlation relationship in order to improve the segmentation results.

6.3. Feature maps visualization

In this section, we illustrate the advantage of our proposed correlation attention module by visualizing the feature representation maps. We select an example to show the feature representation maps of the first layer from four modalities in Figure 7. We denote our proposed method without any fusion strategy as baseline, the first column: input modality, the second column: baseline, third column: 'baseline + dual attention module', the fourth column: 'baseline + tri-attention module (added on the fused feature)', the fifth column: 'baseline + tri-attention module (added on the spatial attention feature)', and the sixth column: ground truth. From Figure 7, we can observe that compared to the baseline, the attention mechanism (column: 3rd, 4th, 5th) allows to highlight feature representations related to brain tumor regions, especially when corre-

lation is taken into account (column: 4th and 5th). In fact, the correlation attention module helps to enhance the fused modality-spatial feature representation for images with fewer information in the tumor region, such as T1 and T2 images.

To further investigate the contribution of the correlation attention module, we used it to guide the fused feature representations (column: 4th) and spatial-attention feature representations (column: 5th), respectively. From Figure 7, we can observe that the correlation attention module added on the spatial-attention feature representations (column: 5th) can further stand out the interested tumor regions for segmentation, and the fuzzy contour becomes clear. We explained that the spatial attention module can help the network to extract the feature representations relating the tumor positions. In conclusion, the correlation attention module can constrain the network to emphasize the interested tumor region for segmentation, revealing that the addition of correlation attention module in the network encourages better segmentation results.

7. Conclusion

In this paper, we proposed a tri-attention fusion guided 3D multi-modal brain tumor segmentation network, where the architecture demonstrated their segmentation performances in multi-modal MR images of glioma patients.

To take advantage of the complimentary information from different modalities, the multi-encoder based network is used to learn modality-specific feature representation. Considering the correlation between MR modalities can help the segmentation, a tri-attention fusion block is proposed, which consists of a modality attention module, a spatial attention module and a correlation attention module. The modality attention module is used to distinguish the contribution of each modality, and the spatial attention module is used to extract more useful spatial information to boost the segmentation result. Since there is a strong correlation between modalities, a correlation description block is used to represent the multi-modal correlation, a correlation based constraint

is introduced to the correlation attention module to guide the network to learn the most correlated feature representation to improve the segmentation. In conclusion, the proposed tri-attention fusion strategy utilized the complimentary information between modalities to encourage the network to learn more useful feature representation to boost the segmentation result.

The advantages of our proposed network architecture (1) The experiment results evaluated on the two metrics (Dice Score and Hausdorff Distance) demonstrate that our proposed method gives an accurate result for the segmentation of brain tumors and its sub-regions even small regions. (2) The architecture are an end-to-end deep leaning approach and fully automatic without any user interventions. (3) The proposed correlation attention module can help the segmentation network to learn correlated feature representations to achieve very competitive results. (4) The proposed correlation attention module can be generalized to other multi-source image processing problem if some correlations exist between them.

However, our work has some limitations that inspire future directions: (1) The work is only validated on multi-modal MR brain tumor images, in the future, we will valid our method in different medical image datasets. (2) The proposed correlation description block is a simple two-layer network, we intend to design a more complex and efficient correlation description block to describe the correlation between multi modalities. (3) The proposed correlation module is applied on brain tumor segmentation, we plan to apply it to synthesize additional images to cope with the limited medical image dataset or deal with the missing modality segmentation issue.

References

[1] Z.-P. Liang, P. C. Lauterbur, Principles of magnetic resonance imaging: a signal processing perspective, SPIE Optical Engineering Press, 2000.

[2] S. Bauer, R. Wiest, L.-P. Nolte, M. Reyes, A survey of mri-based medical

image analysis for brain tumor studies, Physics in Medicine & Biology 58 (13) (2013) R97.

[3] A. Drevelegas, Imaging of brain tumors with histological correlations, Springer Science & Business Media, 2010.

[4] J. Lapuyade-Lahorgue, J.-H. Xue, S. Ruan, Segmenting multi-source images using hidden markov fields with copula-based multivariate statistical distributions, IEEE Transactions on Image Processing 26 (7) (2017) 3187–3195.

[5] T. Zhou, S. Canu, P. Vera, S. Ruan, Brain tumor segmentation with missing modalities via latent multi-source correlation representation, in International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI).

[6] N. Zhang, S. Ruan, S. Lebonvallet, Q. Liao, Y. Zhu, Kernel feature selection to fuse multi-spectral mri images for brain tumor segmentation, Computer Vision and Image Understanding 115 (2) (2011) 256–269.

[7] C. Lian, S. Ruan, T. Denœux, H. Li, P. Vera, Joint tumor segmentation in pet-ct images using co-clustering and fusion based on belief functions, IEEE Transactions on Image Processing 28 (2) (2018) 755–766.

[8] D. Zikic, B. Glocker, E. Konukoglu, A. Criminisi, C. Demiralp, J. Shotton, O. M. Thomas, T. Das, R. Jena, S. J. Price, Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel mr, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2012, pp. 369–376.

[9] Y. Yu, P. Decazes, J. Lapuyade-Lahorgue, I. Gardin, P. Vera, S. Ruan, Semi-automatic lymphoma detection and segmentation using fully conditional random fields, Computerized Medical Imaging and Graphics 70 (2018) 1–7.

[10] S. Bauer, L.-P. Nolte, M. Reyes, Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2011, pp. 354–361.

[11] G. Wang, W. Li, S. Ourselin, T. Vercauteren, Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks, in: International MICCAI Brainlesion Workshop, Springer, 2017, pp. 178–190.

[12] H. Chen, Z. Qin, Y. Ding, L. Tian, Z. Qin, Brain tumor segmentation with deep convolutional symmetric neural network, Neurocomputing 392 (2020) 305–313.

[13] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, Y. Fan, A deep learning model integrating fcnns and crfs for brain tumor segmentation, Medical image analysis 43 (2018) 98–111.

[14] A. Myronenko, 3d mri brain tumor segmentation using autoencoder regularization, in: International MICCAI Brainlesion Workshop, Springer, 2018, pp. 311–320.

[15] J. Wei, Y. Xia, Y. Zhang, M3net: A multi-model, multi-size, and multi-view deep neural network for brain magnetic resonance image segmentation, Pattern Recognition 91 (2019) 366–378.

[16] J. Dolz, C. Desrosiers, L. Wang, J. Yuan, D. Shen, I. B. Ayed, Deep cnn ensembles and suggestive annotations for infant brain mri segmentation, Computerized Medical Imaging and Graphics 79 (2020) 101660.

[17] S. Chen, C. Ding, M. Liu, Dual-force convolutional neural networks for accurate brain tumor segmentation, Pattern Recognition 88 (2019) 90–100.

[18] T. Zhou, S. Ruan, S. Canu, A review: Deep learning for medical image segmentation using multi-modality fusion, Array 3 (2019) 100004.

[19] K. Kamnitsas, W. Bai, E. Ferrante, S. McDonagh, M. Sinclair, N. Pawlowski, M. Rajchl, M. Lee, B. Kainz, D. Rueckert, et al., Ensembles of multiple models and architectures for robust brain tumour segmentation, in: International MICCAI Brainlesion Workshop, Springer, 2017, pp. 450–462.

[20] K.-L. Tseng, Y.-L. Lin, W. Hsu, C.-Y. Huang, Joint sequence learning and cross-modality convolution for 3d biomedical segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6393–6400.

[21] T. Zhou, S. Ruan, Y. Guo, S. Canu, A multi-modality fusion network based on attention mechanism for brain tumor segmentation, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), IEEE, 2020, pp. 377–380.

[22] A. G. Roy, N. Navab, C. Wachinger, Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 421–429.

[23] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, K. H. Maier-Hein, Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge, in: International MICCAI Brainlesion Workshop, Springer, 2017, pp. 287–297.

[24] B. B. Avants, N. Tustison, G. Song, Advanced normalization tools (ants), Insight j 2 (2009) 1–35.

[25] Y. Hu, X. Liu, X. Wen, C. Niu, Y. Xia, Brain tumor segmentation on multimodal mr imaging using multi-level upsampling in decoder, in: International MICCAI Brainlesion Workshop, Springer, 2018, pp. 168–177.

[26] E. Gates, J. G. Pauloski, D. Schellingerhout, D. Fuentes, Glioma segmen-

tation and a simple accurate model for overall survival prediction, in: International MICCAI Brainlesion Workshop, Springer, 2018, pp. 476–484.

[27] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, B. Glocker, Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation, Medical image analysis 36 (2017) 61–78.

[28] T. A. Tuan, et al., Brain tumor segmentation using bit-plane and unet, in: International MICCAI Brainlesion Workshop, Springer, 2018, pp. 466–475.

[29] X. Hu, H. Li, Y. Zhao, C. Dong, B. H. Menze, M. Piraud, Hierarchical multi-class segmentation of glioma images using networks with multi-level activation function, in: International MICCAI Brainlesion Workshop, Springer, 2018, pp. 116–127.

Tongxue Zhou received the M.S.degree in mechatronic engineering in 2017 from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, China. She is currently pursuing the Ph.D. degree at LITIS, INSA Rouen, France. Her current research interests include medical image segmentation, data fusion and deep learning.

Su Ruan received the Ph.D. degrees in "Image Processing" from the University of Rennes 1, France, in 1993. She is currently a full professor with the LITIS Laboratory at the University of Rouen, France. Her research interests include image segmentation and classification, data fusion and pattern recognition.

Pierre Vera received the MD degree in 1993 from Université Paris VI and the PhD degree from the same institution in 1999. He is currently a University Professor and Hospital Physician with the Faculty of Medicine, University of Rouen, France. His research interests include radiation oncology, biophysics, and medical imaging.

Stéphane Canu received the Ph.D.degree in system command from the Compiégne University of Technology in 1986. He received the French Habilitation degree from Paris 6 University. He is currently a Professor of the LITIS

research laboratory at INSA. His research interests includes deep learning, kernels machines and regularization.

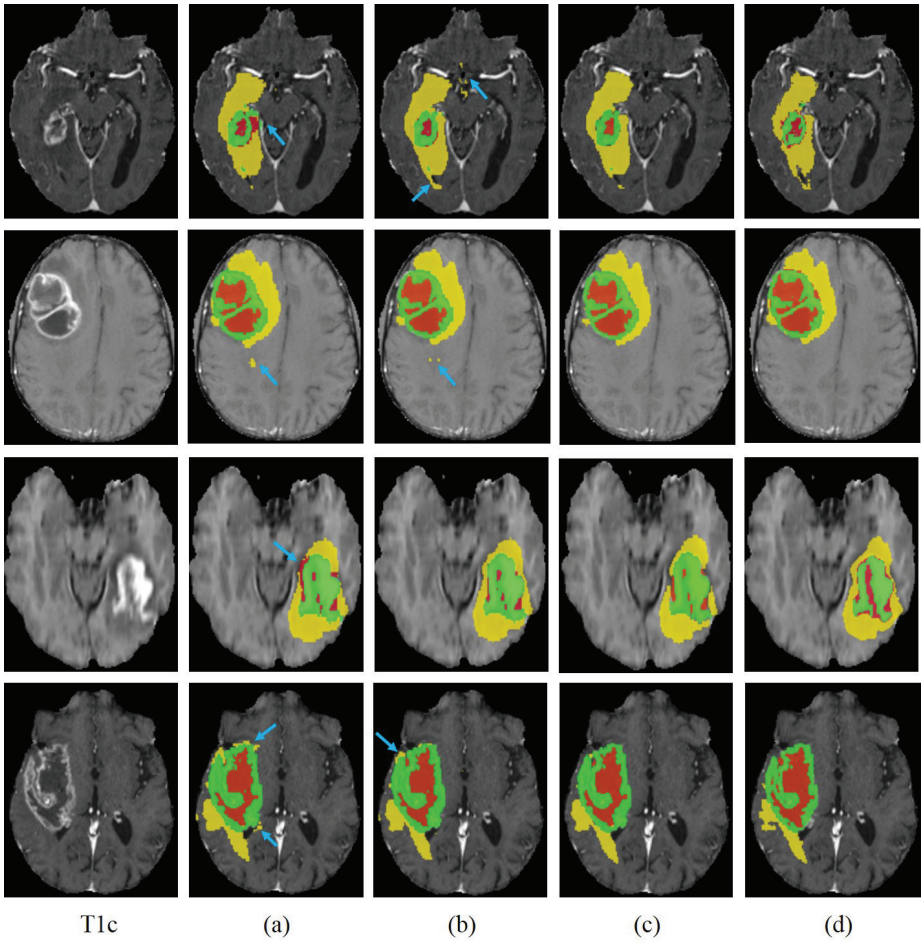

Figure 6: Visualization of several segmentation results. (a) Baseline (b) Baseline with dual attention fusion (c) Baseline with tri-attention fusion (d) Ground truth. Red: necrotic and non-enhancing tumor core; Yellow: edema; Green: enhancing tumor. Blue arrow emphasizes the mis-segmentation of the methods.

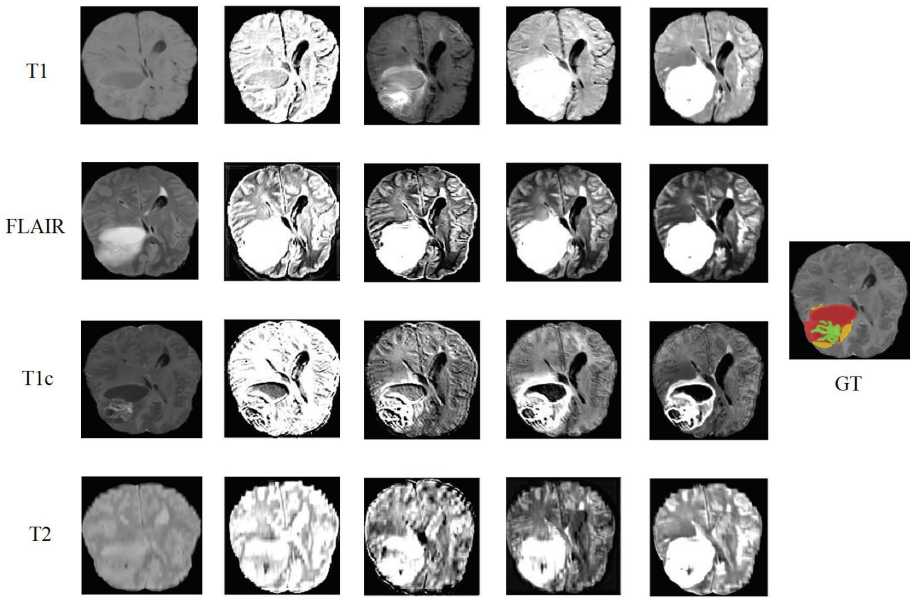

Figure 7: Visualization of effectiveness of proposed correlation attention module. First column: input image, second column: baseline, third column: baseline + dual attention module, fourth column: baseline + tri-attention module (added on fused feature representation), fifth column: (added on spatial-attention feature representation), sixth column: ground-truth.