RFNet: Region-aware Fusion Network for Incomplete Multi-modal Brain Tumor Segmentation

Yuhang Ding[1,2] Xin Yu[2] Yi Yang[2*]
[1] Baidu Research [2] ReLER, University of Technology Sydney
dyh.ustc.uts@gmail.com, {xin.yu, yi.yang}@uts.edu.au

Abstract

Most existing brain tumor segmentation methods usually exploit multi-modal magnetic resonance imaging (MRI) images to achieve high segmentation performance. However, the problem of missing certain modality images often happens in clinical practice, thus leading to severe segmentation performance degradation. In this work, we propose a Region-aware Fusion Network (RFNet) that is able to exploit different combinations of multi-modal data adaptively and effectively for tumor segmentation. Considering different modalities are sensitive to different brain tumor regions, we design a Region-aware Fusion Module (RFM) in RFNet to conduct modal feature fusion from available image modalities according to disparate regions. Benefiting from RFM, RFNet can adaptively segment tumor regions from an incomplete set of multi-modal images by effectively aggregating modal features. Furthermore, we also develop a segmentation-based regularizer to prevent RFNet from the insufficient and unbalanced training caused by the incomplete multi-modal data. Specifically, apart from obtaining segmentation results from fused modal features, we also segment each image modality individually from the corresponding encoded features. In this manner, each modal encoder is forced to learn discriminative features, thus improving the representation ability of the fused features. Remarkably, extensive experiments on BRATS2020, BRATS2018 and BRATS2015 datasets demonstrate that our RFNet outperforms the state-of-the-art significantly.

1. Introduction

Brain tumor segmentation, aiming to segment different brain tumor regions, is vital for clinical assessment and surgical planning. In order to improve the segmentation accuracy, most existing methods [16, 43, 17, 29, 11, 4, 38] use four modalities simultaneously, namely Fluid Attenuation Inversion Recovery (Flair), contrast enhanced T1-weighted (T1c), T1-weighted (T1) and T2-weighted (T2). However, the missing modality problem is very common in clinical practice due to different scanning protocols and patient conditions. Therefore, these standard brain tumor segmentation networks cannot be deployed directly in practice.

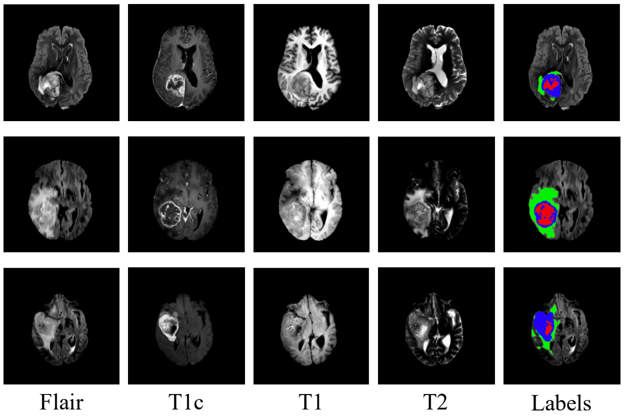

Figure 1. Illustration of different sensitivities of modalities to different brain tumor regions. From left to right: Images of four modalities, *i.e*., Flair, T1c, T1 and T2, and the corresponding labels of three patients are shown. In the segmentation results, different colors denote different brain tumor regions.

Incomplete multi-modal brain tumor segmentation approaches [3, 10, 14, 44] have been proposed to deal with various missing situations. Havaei *et al*. [14] and Dorent *et al*. [10] compute the mean and variance across accessible multi-modal features as fused features. However, this fusion treats each modality equally regardless of different missing scenarios and thus may fail to aggregate features effectively. Later, Chen *et al*. [3] and Zhou *et al*. [44] leverage attention mechanisms to emphasize contributions from different accessible modalities. However, they do not fully exploit the relations between tumor regions and image modalities. In particular, different modalities contain distinct appearances and thus have different sensitivities to diverse tumor regions. For example, as visible in Fig. 1, T1c is more sensitive to the red and blue tumor areas while Flair and T2

[*]This work was done when Yuhang Ding interned at Baidu Research. Yi Yang is the corresponding author.

provide more information for the green tumor area. This observation motivates us that we should pay different attention to different modalities and different regions in order to achieve accurate brain tumor segmentation.

Taking the relations between modalities and regions into account, we propose a Region-aware Fusion Network (RFNet) to aggregate various accessible multi-modal features from different regions adaptively. Our RFNet is constructed by an encoder-decoder architecture, where four encoders are employed to extract features from different modal images. In order to establish the relations between image modalities and tumor regions, we introduce a Region-aware Fusion Module (RFM) into our RFNet. RFM first divides modal features into different regions (*i.e*., tumor sub-structure) via a learned probability map. The probability map indicates the probabilities of tumor regions at each pixel. Then, RFM generates corresponding attention weights in each region to adaptively control the contributions of different image modalities.

Since brain tumors usually occupy a small part of brains, we introduce a region-norm pooling operation to obtain a normalized global feature from each region. Thereby, we prevent the global feature from being numerically too small. Then, we employ two fully-connected layers and a sigmoid activation to attain attention weights from the global feature for image modalities and tumor regions. In this fashion, RFM will generate larger weights for the modalities which are more sensitive to certain tumor regions, thus leading to discriminative fused features for accurate segmentation.

Due to the missing hetero-modal data, RFNet will face the problem of unbalanced training. To be specific, RFNet might try to seek the easiest way to segment brain tumors from the multi-modal data. In other words, the network segments each region mainly by exploiting the modalities which are sensitive to the region rather than all the modality information. However, this will lead to poor segmentation accuracy when some modalities are missing. To tackle this problem, we develop a segmentation-based regularizer. In particular, a weight-shared decoder is employed to segment each modality individually. In this manner, each modal encoder is forced to learn discriminative features for all the tumor regions. Therefore, RFNet can segment different regions well even when some modalities are missing. Benefit from the proposed fusion module and regularizer, RFNet achieves higher accuracy than the state-of-the-art methods on BRATS2020, BRATS2018 and BRATS2015. This demonstrates the superiority of our method.

Overall, our contributions are threefold:

- We propose a Region-aware Fusion Network (RFNet) for incomplete multi-modal brain tumor segmentation. Particularly, we introduce a novel a Region-aware Fusion Module (RFM) by explicitly taking the relations between modalities and regions into account. With the help of RFM, RFNet effectively aggregates diverse combinations of modal features and produces discriminative fused features for segmentation.
- To address the unbalanced training problem of RFNet, we propose a segmentation-based regularizer. The proposed regularizer enforces each modal encoder to produce discriminative features for segmenting all the tumor regions, thus further improving the discriminativeness of the fused features.
- Taking advantage of the proposed fusion module and regularizer, RFNet achieves superior segmentation accuracy compared to the state-of-the-art on the widely-used BRATS2020, BRATS2018 and BRATS2015 benchmarks.

2. Related Work

In this section, we briefly review the most related works on incomplete multi-modal brain tumor segmentation. Moreover, as we propose a feature fusion module to tackle the missing modality problem, existing deep multi-modal fusion methods are also reviewed.

2.1. Incomplete Multi-modal Tumor Segmentation

Incomplete data is very common in practical applications, such as scarce annotation problems [40, 31, 8] and missing modality problems [37, 28, 19, 42]. In this work, we focus on incomplete multi-modal brain tumor segmentation, which aims to segment brain tumors from various missing hetero-modal MRI images. Therefore, compared with the standard brain tumor segmentation [13, 21, 41, 23], segmenting brain tumors from incomplete multi-modal data is more practical but more challenging.

Shen *et al*. [24] treat various missing modalities as different domains and then leverage adversarial learning to project images from these domains into a unified feature space during segmentation. However, since it is difficult to align distinct and diverse distributions simultaneously, their method can only handle a small number of missing modalities. Zhou *et al*. [44] generate the features of missing modalities according to the correlations between different modalities. However, their method may be not suitable when few modalities are available, because only one or two modalities are not enough to generate reliable features for the missing modalities.

In addition to feature alignment [24] and feature completion [44], several prior works [10, 3, 14] attempt to leverage feature fusion to solve the missing modality problem: Havaei *et al*. [14] aggregate partial modalities by calculating the mean and variance of the available features. Dorent *et al*. [10] embed all observed modalities into a shared latent representation by employing a multi-modal variational auto-encoder. Chen *et al*. [3] aggregate incomplete modalities via concatenation and leverage feature disentanglement jointly to obtain a modality-invariant and discriminative representation. However, these prior arts all do not fully consider the relations between brain tumor regions and different modalities and thus do not aggregate features effectively. In contrast, our RFNet fuses features in a region-aware manner and thus obtains discriminative information for each region segmentation. Besides, we also propose a segmentation-based regularizer to facilitate the training process of RFNet.

2.2. Deep Multi-modal Fusion

In recent years, the proliferation of multi-modal applications [36, 12, 2, 30, 1] has been witnessed, such as multi-view classification [12, 39], multi-view localization [25, 26, 27], visual question answering [2] and visual language navigation [1, 33, 5] Accordingly, effective multi-modal fusion techniques [34, 22, 9] have also received substantial research attention.

Wang *et al*. [34] propose a channel-exchanging network to aggregate modalities in a parameter-free manner, while Perez-Rua *et al*. [22] adopt the architecture search to design an optimal feature fusion module for a given dataset. Dolz *et al*. [9] introduce dense connections not only in each modal network but also between two modal embedding networks for feature fusion. Unlike previous work, we explicitly take the relationships between image modalities and tumor regions/sub-structure into account and then design a fusion module based on this observation. Thus, our RFM is able to fuse different modal features adaptively and thus generates more discriminative features for segmentation.

3. Proposed Method

In this work, we design RFNet for incomplete multi-modal brain tumor segmentation. In particular, we develop an RFM module to take advantage of available modalities effectively during feature fusion. In addition, we also propose a segmentation-based regularizer to further improve the feature representations of each modal encoder, thus facilitating the final segmentation performance. In this section, we will introduce our designed RFNet as well as the proposed regularization term in detail.

3.1. Task Definition

Incomplete multi-modal brain tumor segmentation aims to segment three brain tumor areas, *i.e*., the whole tumor, the tumor core and the enhancing tumor, from various combinations of multi-modal MRI images, including Flair, T1c, T1 and T2. The whole tumor is composed of all the three tumor sub-regions, *i.e*., the necrotic and non-enhancing tumor core (NCR/NET), the peritumoral edema (ED), and the GD-enhancing tumor (ET). The tumor core consists of NCR/NET and ET while the enhancing tumor involves ET. Figure 1 illustrates NCR/NET, ED and ET in red, green and blue, respectively.

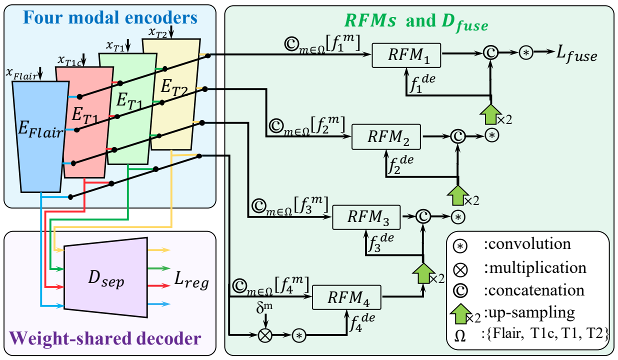

Figure 2. Illustration of our proposed RFNet. Four encoders, *i.e*., $\mathbf{E}_{\text{Flair}}$, $\mathbf{E}_{\text{T1c}}$, $\mathbf{E}_{\text{T1}}$ and $\mathbf{E}_{\text{T2}}$, are employed to extract features from four modalities individually. $\mathbf{D}_{\text{sep}}$ is our segmentation-based regularizer network, while $\mathbf{D}_{\text{fuse}}$ with the designed RFM is used to attain the final segmentation predictions. δ^m simulates different missing scenarios.

In order to measure the robustness of our proposed method against various missing scenarios, we evaluate its segmentation results on all the fifteen combinations of the four image modalities and the average score is reported for comparisons. During training, all modalities and labels are available and we simulate missing scenarios by setting missing modal features to zero.

3.2. Architecture Overview

We adopt a 3D U-Net [6] architecture with a late fusion strategy to construct our RFNet. As shown in Fig. 2, four encoders, *i.e*., $\{\mathbf{E}_m\}_{m\in\{\text{Flair},\text{T1c},\text{T1},\text{T2}\}}$, are employed to extract features from four modalities separately. The decoder $\mathbf{D}_{\text{sep}}$ is designed to segment each modality separately, thus assisting our four encoders in learning representative region features. Furthermore, $\mathbf{D}_{\text{sep}}$ shares weights for the four image modalities, so that four modal features can be projected into a shared latent space. This also significantly facilitates the later feature aggregation and fusion.

The decoder $\mathbf{D}_{\text{fuse}}$ is designed to obtain the final segmentation results from the aggregated features, as visible in Fig. 2. In each stage, the encoder features are fused by the designed RFM. Note that, RFM takes not only four encoder features but also the features from the prior layer as input. This is because that the previous layer features can be used to embed semantic information of tumor regions, thus making RFM region-aware. In the bottleneck (*i.e*., the fourth stage S_4), there are no previous layer features available for RFM. Therefore, we leverage an additional convolutional layer to embed the encoder features into semantic features for the fusion module in Fig. 2.

3.3. Region-aware Fusion Module

Considering different sensitivities of image modalities to different regions, as shown in Fig. 1, our RFNet aims to pay different attention to different modalities in each region. In this fashion, discriminative features for tumor regions can be obtained, leading to the improvement of segmentation accuracy. To this end, we develop an RFM module that is designed to fuse available modal features in a region-aware fashion, as visible in Fig. 3. RFM mainly consists of two parts: probability map learning and region-aware multi-modal feature fusion.

Probability Map Learning: To achieve the region-aware characteristics, our RFM first learns a probability map which indicates the probabilities of brain tumor structure (including healthy brain regions) at each location. As shown in Fig. 3, the probability map is obtained from the decoder feature of the previous layer f^{de} and the available encoder features $©_{m\in\Omega}\,[f^{m}\cdot\delta^{m}]$. Employing the encoder features in RFM is because that they offer more detailed spatial information and can improve the accuracy of the probability maps. © denotes the concatenation operation while Ω denotes the modality set, including Flair, T1c, T1 and T2. δ^m is set to either 0 or 1, indicating whether the m modality is missing or not. The probability map learning procedure is defined as:

$$\hat{y}^{pm}_{i,j} = \frac{\exp(\phi_j(f^{pm}_{i,j};\theta_j))}{\sum_{k\in K}\exp(\phi_j(f^{pm}_{i,j};\theta_j)_k)}, \tag{1}$$

where $f^{pm}_{i,j}$ represents the features from $f^{de}_{i,j}$ and $©_{m\in\Omega}\left[f^{m}_{i,j}\cdot\delta^{m}\right]$. i and j denote the i-th subject and the j-th stage/level of the network, respectively. $\hat{y}^{pm}_{i,j}$ is the learned probability map. ϕ_j denotes the region classifier in the j-th stage and θ_j is the corresponding parameters. K denotes the brain tumor region set, including BG (background), NCR/NET, ED and ET.

The probability map (shown in Fig. 4) is learned under the supervision of the ground truth by a weighted cross-entropy loss $\mathcal{L}_{WCE}$ [3] and a Dice loss $\mathcal{L}_{DL}$, expressed as:

$$\mathcal{L}_{pm} = \sum_{i=1}^{N}\sum_{j=1}^{S_{num}}\big(\mathcal{L}_{WCE}(\psi_j(\hat{y}^{pm}_{i,j}),y_i)+\mathcal{L}_{DL}(\psi_j(\hat{y}^{pm}_{i,j}),y_i)\big), \tag{2}$$

where N and S_{num} denote the number of training data and stages. ψ_j denotes the up-sampling operation in the j-th stage, aiming to match the resolution of the probability map $\hat{y}^{pm}_{i,j}$ and the ground-truth mask y_i. $\mathcal{L}_{WCE}$ is formulated as:

$$\mathcal{L}_{WCE}(\hat{y},y) = \sum_{k\in K}\frac{\|-\alpha_k\cdot y_k\cdot\log(\hat{y}_k)\|_1}{H\cdot W\cdot Z}, \tag{3}$$

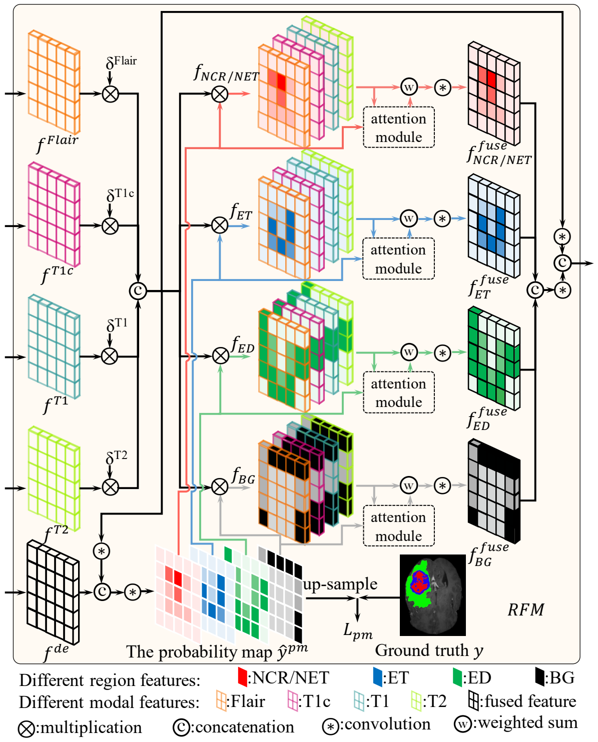

Figure 3. Illustration of our region-aware fusion module (RFM). The probability map is firstly learned to divide multi-modal features into different regions. Then, an attention mechanism is designed to aggregate features in a region-aware manner.

where $\|\cdot\|_1$ denotes the L1 norm, and W, H and Z denote the width, height and slice number of the 3D volumes, respectively. α_k is the weight for the region k and $\alpha_k = 1-\frac{\|y_k\|_1}{\sum_{k'\in K}\|y_{k'}\|_1}$. $\mathcal{L}_{DL}$ is formulated as:

$$\mathcal{L}_{DL}(\hat{y},y) = 1-\sum_{k\in K}\frac{2\cdot\|\hat{y}_k\bigcap y_k\|_1}{K_{num}\cdot(\|\hat{y}_k\|_1+\|y_k\|_1)}, \tag{4}$$

where $\bigcap$ denotes the overlap between predictions and ground-truth masks, and K_{num} denotes the number of regions in K.

Region-aware Multi-modal Fusion: With the help of the probability map, RFM has managed to divide multi-modal features into different regions. Thus, the region-aware fusion is conducted on the divided features in each region.

The feature division is implemented by multiplying features with the probability map, written as:

$$f_k = ©_{m\in\Omega}\,[f^{m}\cdot\delta^{m}]\cdot\hat{y}^{pm}_{k}, \tag{5}$$

where f_k[1] denotes the divided features of the available modalities in the tumor region k and f^m denotes the encoder feature of the modality m.

As shown in Fig. 3, after feature division, modal-wise attention weights are learned individually in different regions

[1] For simplicity, we omit the subscripts i and j without causing any confusion.

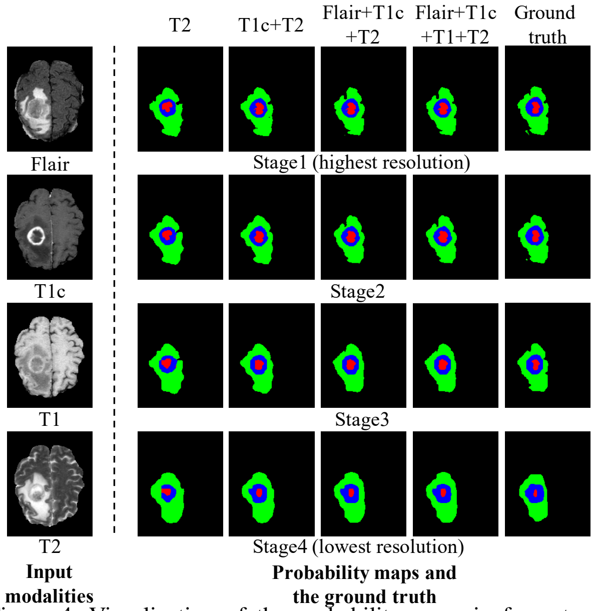
T2
T1c+T2
Flair+T1c +T2
Flair+T1c +T1+T2
Ground truth
Flair
Stage1 (highest resolution)
T1c
Stage2
T1
Stage3
T2
Stage4 (lowest resolution)
Input modalities
Probability maps and the ground truth

Figure 4. Visualization of the probability maps in four stages. Left: four image modalities. Right: Estimated probability maps from different combinations of image modalities in different stages/levels of our network and the corresponding ground truth.

to aggregate the corresponding features. Figure 5 illustrates the generation procedure of the attention weight in the region k. Specifically, the global feature of the region k is obtained via an average pooling operation and is then normalized by the probability map $\hat{y}_k^{pm}$. Employing this region-norm pooling can prevent the averaged global feature from being numerically too small, given the fact that brain tumors usually occupy only a small area in a brain. Then, two fully-connected layers, along with a Leaky ReLU layer and a sigmoid activation, are adopted to embed the normalized feature into a modal-wise attention weight. As shown in Fig 5, the generated attention weights are then applied to the divided feature f_k to adjust the contributions from available modalities to obtain discriminative fused features.

Considering the distinct sensitivities of different modalities in various regions, RFM employs separate attention modules for each region to generate corresponding attention weights, as shown in Fig. 3. By paying larger attention to more sensitive modalities, RFM is able to generate more representative features for each region. To feed these region features to the decoder, in Fig. 3, RFM adopts a concatenation operation followed by a convolutional bottleneck. A shortcut connection is also employed, similar to the residual learning [15].

3.4. Segmentation-based Regularizer

The phenomenon of missing multi-modal data usually introduces unbalanced training issues [32]. To be specific, deep neural networks usually opt to segment tumor regions mainly based on the discriminative modalities. Therefore, some modal encoders are well trained to be able to identify the corresponding tumor regions while other encoders are not. This would lead to severe accuracy degradation in tumor segmentation when the discriminative modalities are missing.

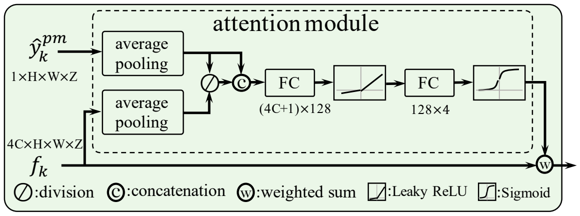

Figure 5. Illustration of the attention module. The region-norm pooling normalizes the global feature of f_k by the average probability of $\hat{y}_k$ to obtain the features to generate the attention weights.

To solve this problem, we propose a segmentation-based regularizer. As illustrated in Fig. 2, RFNet adopts a weight-shared decoder $\mathbf{D}_{\text{sep}}$ to segment each modal image separately. The corresponding weighted cross-entropy loss and Dice loss are employed as the regularization term, written as:

$$\mathcal{L}_{reg} = \sum_{i=1}^{N} \sum_{m \in \Omega} \left(\mathcal{L}_{WCE}(\hat{y}_{i,m}^{sep}, y_i) + \mathcal{L}_{DL}(\hat{y}_{i,m}^{sep}, y_i) \right), \quad (6)$$

where $\hat{y}_{i,m}^{sep}$ denotes the predicted segmentation mask of the i-th subject from the modality m. The segmentation-based regularizer enforces each modal encoder to be discriminative to each tumor region. In this manner, RFNet is able to obtain representative encoder features, thus improving the segmentation performance.

3.5. Overall Loss

As shown in Fig. 2, $\mathbf{D}_{\text{fuse}}$ is employed to predict the final segmentation mask from the fused features. The weighted cross-entropy loss and Dice loss are used to align the predictions to the corresponding ground-truth segmentation maps, expressed as:

$$\mathcal{L}_{fuse} = \sum_{i=1}^{N} \left(\mathcal{L}_{WCE}(\hat{y}_i^{fuse}, y_i) + \mathcal{L}_{DL}(\hat{y}_i^{fuse}, y_i) \right), \quad (7)$$

where $\hat{y}_i^{fuse}$ is the predicted segmentation mask from the i-th subject. Therefore, the overall loss of our RFNet is defined as:

$$\mathcal{L} = \mathcal{L}_{pm} + \mathcal{L}_{reg} + \mathcal{L}_{fuse}. \quad (8)$$

4. Experiments

4.1. Implementation Details

RFNet adopts 3D-Unet [6] with four-stage encoders ($\{\mathbf{E}_m\}_{m \in \Omega}$) and decoders ($\mathbf{D}_{\text{sep}}$ and $\mathbf{D}_{\text{fuse}}$). The architecture details can be referred to the supplementary material.

Table 1. Comparisons with the state-of-the-art methods, including HeMIS [14], U-HVED [10] and RobustSeg [3], on BRATS2020. Complete, Core and Enhancing denote the Dice scores of the whole tumor, the tumor core and the enhancing tumor, respectively. All the results are reproduced by using the authors' codes.

Modalities				Dice scores (%)											
				Complete				Core				Enhancing			
F	T1	T1c	T2	HeMIS	U-HVED	RobustSeg	Ours	HeMIS	U-HVED	RobustSeg	Ours	HeMIS	U-HVED	RobustSeg	Ours
○	○	○	●	79.85	80.75	82.20	**86.05**	54.22	57.43	61.88	**71.02**	31.43	28.70	36.46	**46.29**
○	○	●	○	64.58	68.54	71.39	**76.77**	69.41	73.01	76.68	**81.51**	63.24	66.59	67.91	**74.85**
○	●	○	○	63.01	54.93	71.41	**77.16**	42.42	36.73	54.30	**66.02**	16.53	12.33	28.99	**37.30**
●	○	○	○	52.29	82.69	82.87	**87.32**	24.97	51.15	60.72	**69.19**	9.00	20.87	34.68	**38.15**
○	○	●	●	84.45	83.37	85.97	**87.74**	77.60	77.85	82.44	**83.45**	70.30	68.74	71.42	**75.93**
○	●	●	○	72.50	71.58	76.84	**81.12**	75.59	76.49	80.28	**83.40**	70.71	67.82	70.11	**78.01**
●	●	○	○	65.29	85.01	88.10	**89.73**	41.58	55.10	68.18	**73.07**	13.99	22.53	39.67	**40.98**
○	●	○	●	82.31	81.58	85.53	**87.73**	56.38	59.29	66.46	**73.13**	28.58	28.73	39.92	**45.65**
●	○	○	●	81.56	87.40	88.09	**89.87**	55.89	61.87	68.20	**74.14**	28.91	30.48	42.19	**49.32**
●	○	●	○	69.37	86.13	87.33	**89.89**	70.86	76.86	81.85	**84.65**	68.31	69.53	70.78	**76.67**
●	●	●	○	73.31	87.10	88.87	**90.69**	75.07	79.51	82.76	**85.07**	70.80	71.32	71.77	**76.81**
●	●	○	●	83.03	88.07	89.24	**90.60**	57.40	63.46	70.46	**75.19**	29.53	30.60	43.90	**49.92**
●	○	●	●	84.64	88.33	88.68	**90.68**	77.69	78.68	81.89	**84.97**	71.36	69.84	71.17	**77.12**
○	●	●	●	85.19	84.27	86.63	**88.25**	79.05	79.99	82.85	**83.47**	71.67	69.74	71.87	**76.99**
●	●	●	●	85.19	88.81	89.47	**91.11**	78.58	80.40	82.87	**85.21**	71.49	70.50	71.52	**78.00**
Average				75.10	81.24	84.17	**86.98**	65.45	67.19	73.45	**78.23**	47.73	48.55	55.49	**61.47**

For the image pre-processing, the MRI images are skull-stripped, co-registered and re-sampled to $1mm^3$ resolution by the data collector. In this work, following [3, 10], we additionally cut out the black background area outside the brain and normalize each MRI modality to zero mean and unit variance in the brain area.

During training, input images are randomly cropped to $80\times 80\times 80$ and are then augmented with random rotations, intensity shifts and mirror flipping. We train our network for 300 epochs with the batch size of 2. Adam [18] is leveraged to optimize the network with β_1 and β_2 of 0.9 and 0.999 respectively, and the weight decay is set to $1e^{-5}$. Besides, we adopt the "poly" learning rate policy where the initial learning rate $2e^{-4}$ is multiplied by $(1-\frac{epoch}{max\_epoch})^p$ with $p=0.9$.

Following [3], we firstly segment the $80\times 80\times 80$ patches which slide on the test images and have 50% overlaps over the neighboring patches. Then, the final segmentation map is obtained by fusing the predictions of these patches. Since not all brain tumors contain enhancing areas, we employ a post-processing strategy to reduce the false alarm of enhancing tumors. To be specific, when the number of the pixels predicted as the enhancing tumor is too small (*i.e.*, less than 500), we believe this is a false alarm and we will treat these pixels as non-enhancing tumors as in [10].

4.2. Datasets and Evaluation Metric

Datasets: We evaluate RFNet on three datasets from Multimodal Brain Tumor Segmentation Challenge (BRATS) [20], *i.e.*, BRATS2020, BRATS2018 and BRATS2015. The subjects in the three datasets all contain four distinct MRI modalities, *i.e.*, Flair, T1c, T1 and T2.

BRATS2020 contains 369 training subjects which are randomly split by us into 219, 50 and 100 subjects for training, validation and test, respectively. **BRATS2018** contains 285 training subjects which are split into 199, 29 and 57 subjects for training, validation and test, respectively. Besides, we use a three-fold validation with the same split lists as [10] in BRATS2018. **BTATS2015** contains 274 training subjects. Following [14, 3], we divide the dataset into 242, 12 and 20 subjects for training, validation and test, respectively. Since BRATS2020 is the newest and largest dataset, in this work, we mainly focus on BRATS2020.

Evaluation Metric: Dice coefficient [7] is used to measure the segmentation performance of the proposed method, defined as:

$$\mathrm{Dice}_{\bar{k}}(\hat{y}, y) = \frac{2\cdot\|\hat{y}_{\bar{k}} \bigcap y_{\bar{k}}\|_1}{\|\hat{y}_{\bar{k}}\|_1 + \|y_{\bar{k}}\|_1}, \tag{9}$$

Where $\bar{k}$ denotes different tumor classes, including the whole tumor, the tumor core and the enhancing tumor. $\mathrm{Dice}_{\bar{k}}$ denotes the Dice score of the tumor class $\bar{k}$. Larger Dice scores represent that predictions are more similar to the ground truth, and thus indicate better segmentation accuracy.

4.3. Comparisons with the State-of-the-art

In Table 1 and Fig. 6, we compare our RFNet with three state-of-the-art methods, including HeMIS [14], U-HVED [10] and RobustSeg [3]. HeMIS [14] leverages the mean and variance of available modal features as the aggregated feature for segmentation. U-HVED [10] introduces multi-modal variational auto-encoders (MVAE) [35] to project different incomplete multi-modal images into a shared la-

Table 2. Ablation study on RFNet. The average Dice scores of fifteen multi-modal combinations are reported. "Reg": the proposed segmentation-based regularizer, "RFM": the developed region-aware fusion module, "PostPro": the post-processing technique.

Methods	Average Dice scores (%)		
	Complete	Core	Enhancing
Baseline	83.20	71.72	53.73
+RFM	85.07	75.91	56.78
+Reg	86.07	76.89	57.96
+Reg+RFM	**86.98**	**78.23**	59.05
+Reg+RFM+PostPro	**86.98**	**78.23**	**61.47**

Table 3. The necessity of our regularizer and RFM. "wi rec regularizer': employing a reconstruction-based regularizer rather than the segmentation-based regularizer. "modal-wise" and "channel-wise": applying modal-wise and channel-wise attention to the feature maps instead of in a region-aware manner.

Methods	Average Dice scores (%)		
	Complete	Core	Enhancing
wi rec regularizer	85.38	75.50	59.64
channel-wise	85.81	76.36	60.11
modal-wise	85.87	77.02	61.01
RFNet	**86.98**	**78.23**	**61.47**

tent space. RobustSeg [3] disentangles content codes from appearance ones for segmentation and introduces a gated feature fusion to aggregate multi-modal features. These methods all do not explicitly take advantage of the relations between modalities and regions, and neglect the unbalanced training problem.

As shown in Table 1, our method achieves superior segmentation performance. For example, compared with the second best method, *i.e.*, RobustSeg [3], our RFNet improves the average Dice scores by 2.81%, 4.78% and 5.98% in the whole tumor, the tumor core and the enhancing tumor, respectively. Moreover, our method outperforms the state-of-the-art methods on all fifteen multi-modal combinations. This demonstrates the superiority of our method.

4.4. Ablation Study

In Table 2, we conduct ablation study on RFNet. The baseline model leverages a $3 \times 3 \times 3$ convolutional layer to aggregate encoder features. As shown in Table 2, the proposed region-aware fusion module and the segmentation-based regularizer can both improve the network significantly. For example, employing RFM increases the average Dice scores of three tumor areas by 1.87%, 4.19% and 3.05%, respectively. This is because RFM manages to effectively aggregate features and thus provides representative information for segmentation. Moreover, since the proposed regularizer helps the modal encoders to be discriminative to each region, applying the regularizer with RFM further improves segmentation results, as visible in Table 2. The post-processing technique is introduced to reduce false alarms of enhancing tumors, thus improving the segmentation performance of enhancing tumors.

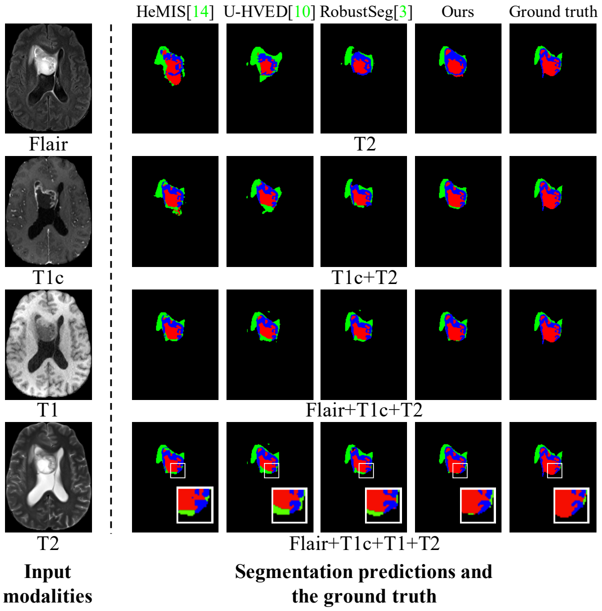

Figure 6. Comparisons with the state-of-the-art. Left: four image modalities. Right: segmentation masks predicted by different methods from different combinations of image modalities and the corresponding ground truth.

To demonstrate the effectiveness of the region-aware operation, we apply a modal-wise attention to each modal feature (*i.e.*, a scalar for each modality) and a channel-wise attention to all the concatenated features. As shown in Table 3, the model without the proposed region-aware operation yields inferior segmentation accuracy. This is because that applying the same attention weights, either modal-wise or channel-wise attention, to the entire image does not enable a network to focus on the tumor regions. In Table 3, a reconstruction-based regularizer is adopted to replace the proposed segmentation-based regularizer and achieves inferior performance. This is because the reconstruction-based regularizer mainly focuses on restoring brain appearances rather than learning discriminative representations for tumor segmentation.

4.5. Comparisons in BRATS2015 and BRATS2018

In addition to BRATS2020, we also compare our method with the state-of-the-art on BRATS2015 and BRATS2018 in Table 4 and Table 5, respectively. Note that, U-HVED [10] and RobustSeg [3] conduct experiments on only one dataset, *e.g.*, BRATS2018 or BRATS2015. Therefore, we obtain the BRATS2015 accuracy of U-HVED [10] with their official code and attain the BRATS2018 results of RobustSeg [3] from the authors. As shown in Table 4 and Table 5, our method improves the segmentation accuracy significantly on both two datasets. For instance, the average Dice scores of the three tumor areas on BRATS2018 are boosted

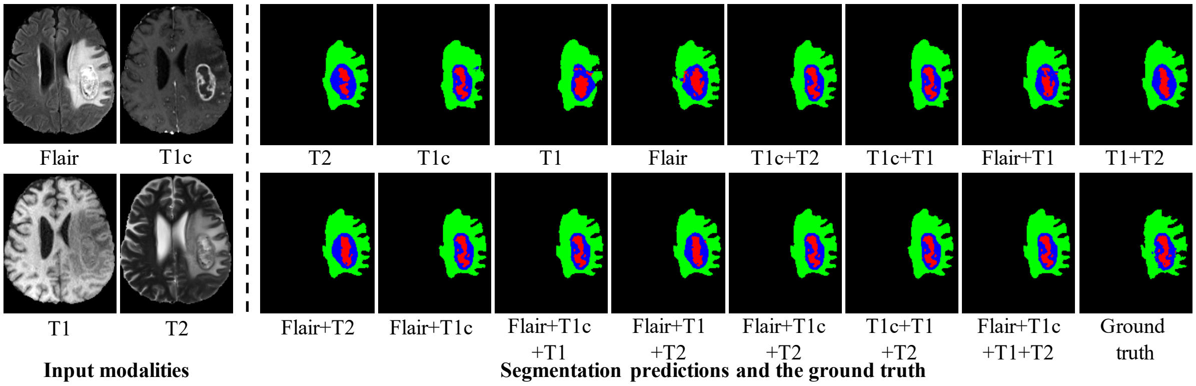

Figure 7. Visualization of the predicted segmentation maps. Left: four image modalities. Right: segmentation maps predicted by our RFNet from all fifteen combinations of image modalities and the corresponding ground truth. More visual results are provided in the supplementary materials.

Table 4. Comparisons with the state-of-the-art on BRATS2015. "†": reproduced based on the authors' code.

Methods	Average Dice scores (%)		
	Complete	Core	Enhancing
HeMIS [14]	68.22	54.07	43.86
U-HVED† [10]	81.57	64.68	56.76
RobustSeg [3]	84.45	69.19	57.33
Ours	**86.13**	**71.93**	58.98
Ours+PostPro	**86.13**	**71.93**	**64.13**

Table 5. Comparisons with the state-of-the-art on BRATS2018. "*": provided by the authors.

Methods	Average Dice scores (%)		
	Complete	Core	Enhancing
HeMIS [14]	78.60	59.70	48.10
U-HVED [10]	80.10	64.00	50.00
RobustSeg* [3]	84.37	69.78	51.02
Ours	**85.67**	**76.53**	54.15
Ours+PostPro	**85.67**	**76.53**	**57.12**

by 1.30%, 6.75% and 6.10% by our RFNet. This validates the superiority of our method.

4.6. Visualization

Visualization of the Segmentation Results: In Fig. 7, we visualize the segmentation results of RFNet from all fifteen multi-modal combinations. Figure 7 illustrates that our method is able to segment brain tumors well in various missing scenarios. For example, RFNet predicts an accurate segmentation map with only Flair and T1c modal images.

Visualization of the Attention Weights: In Fig. 8, we illustrate our generated attention weights which are employed to fuse available modalities adaptively in each region. Since the deeper stage in RFNet encodes high-level semantic information which is vital for segmentation, we opt to visualize the attention weights at the fourth stage. More examples can be refer to supplementary materials. During inference, since missing modal features (zero tensors) provides no information, we set the corresponding attention weights to zero. As shown in Fig. 8, T1c modality (in red) receives more attention in NCR/NET and ET, while in ED, more attention is paid to Flair (in blue) and T2 (in yellow) modalities. This is consistent with the observation in Fig. 1, where T1c is more sensitive to NCR/NET and ET while Flair and T2 are sensitive to ED. Therefore, RFNet is able to provide larger attention weights for the sensitive modalities and thus obtains discriminative features for each region.

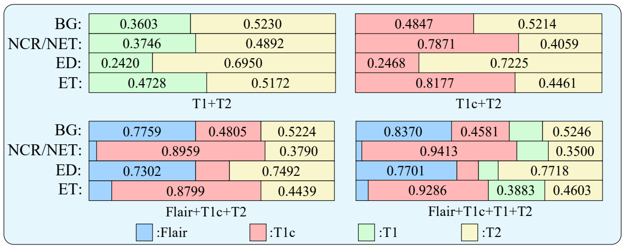

Figure 8. Visualization of the generated attention weights by our RFM at the fourth stage. The four panels demonstrate different cases of missing modalities. In each panel, attention weights (in numbers) are used to aggregate available modalities (in colors) adaptively in diverse regions (in rows). Larger colored boxes denote larger attention weights for the corresponding modality.

5. Conclusion

In this paper, we propose a region-aware fusion network (RFNet) to aggregate various available modalities effectively for incomplete multi-modal brain tumor segmentation. Our newly designed region-aware fusion module (RFM) takes the fact that different modalities exhibit distinct sensitivities to brain tumor regions into account. Thus, RFM achieves more representative fused features from different modal images for accurate segmentation. Moreover, our developed segmentation-based regularizer not only improves the feature representations extracted by our modal encoders in each tumor region but also expedites the training of our RFNet. Extensive experiments demonstrate that our method significantly outperforms the state-of-the-art.

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[3] Cheng Chen, Qi Dou, Yueming Jin, Hao Chen, Jing Qin, and Pheng-Ann Heng. Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 447–456, 2019.

[4] Chen Chen, Xiaopeng Liu, Meng Ding, Junfeng Zheng, and Jiangyun Li. 3d dilated multi-fiber network for real-time brain tumor segmentation in mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 184–192. Springer, 2019.

[5] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547, 2019.

[6] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 424–432, 2016.

[7] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[8] Yuhang Ding, Xin Yu, and Yi Yang. Modeling the probabilistic distribution of unlabeled data for one-shot medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[9] Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed. Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation. *IEEE Transactions on Medical Imaging*, 38(5):1116–1126, 2018.

[10] Reuben Dorent, Samuel Joutard, Marc Modat, Sébastien Ourselin, and Tom Vercauteren. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 74–82, 2019.

[11] Lucas Fidon, Wenqi Li, Luis C Garcia-Peraza-Herrera, Jinendra Ekanayake, Neil Kitchen, Sebastien Ourselin, and Tom Vercauteren. Scalable multimodal convolutional networks for brain tumour segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 285–293. Springer, 2017.

[12] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. *International Conference on Learning Representations*, 2021.

[13] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017.

[14] Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. Hemis: Hetero-modal image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 469–477, 2016.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[16] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[17] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, 2017.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014.

[19] Yanbei Liu, Lianxi Fan, Changqing Zhang, Tao Zhou, Zhitao Xiao, Lei Geng, and Dinggang Shen. Incomplete multi-modal representation learning for alzheimer's disease diagnosis. *Medical Image Analysis*, 69:101953, 2021.

[20] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.

[21] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, pages 311–320. Springer, 2018.

[22] Juan-Manuel Pérez-Rúa, Valentin Vielzeuf, Stéphane Pateux, Moez Baccouche, and Frédéric Jurie. Mfas: Multimodal fusion architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6966–6975, 2019.

[23] Haocheng Shen, Ruixuan Wang, Jianguo Zhang, and Stephen J McKenna. Boundary-aware fully convolutional network for brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 433–441. Springer, 2017.

[24] Yan Shen and Mingchen Gao. Brain tumor segmentation on mri with missing modalities. In *International Conference on Information Processing in Medical Imaging*, pages 417–428, 2019.

[25] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32:10090–10100, 2019.

[26] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4064–4072, 2020.

[27] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11990–11997, 2020.

[28] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1405–1414, 2017.

[29] Kuan-Lun Tseng, Yen-Liang Lin, Winston Hsu, and Chung-Yang Huang. Joint sequence learning and cross-modality convolution for 3d biomedical segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6393–6400, 2017.

[30] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3343–3352, 2019.

[31] Shuxin Wang, Shilei Cao, Dong Wei, Renzhen Wang, Kai Ma, Liansheng Wang, Deyu Meng, and Yefeng Zheng. Lt-net: Label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9162–9171, 2020.

[32] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020.

[33] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6629–6638, 2019.

[34] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *Advances in Neural Information Processing Systems*, 33, 2020.

[35] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, 2018.

[36] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6292–6300, 2019.

[37] Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. Deep partial multi-view learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[38] Dingwen Zhang, Guohai Huang, Qiang Zhang, Jungong Han, Junwei Han, Yizhou Wang, and Yizhou Yu. Exploring task structure for brain tumor segmentation from multi-modality mr images. *IEEE Transactions on Image Processing*, 29:9032–9043, 2020.

[39] Shunli Zhang, Xin Yu, Yao Sui, Sicong Zhao, and Li Zhang. Object tracking with multi-view support vector machines. *IEEE Transactions on Multimedia*, 17(3):265–278, 2015.

[40] Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8543–8553, 2019.

[41] Xiaomei Zhao, Yihong Wu, Guidong Song, Zhenye Li, Yazhuo Zhang, and Yong Fan. A deep learning model integrating fcnns and crfs for brain tumor segmentation. *Medical Image Analysis*, 43:98–111, 2018.

[42] Tan Zhi-Xuan, Harold Soh, and Desmond Ong. Factorized inference in deep markov models for incomplete multimodal time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10334–10341, 2020.

[43] Chenhong Zhou, Changxing Ding, Zhentai Lu, Xinchao Wang, and Dacheng Tao. One-pass multi-task convolutional neural networks for efficient brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 637–645. Springer, 2018.

[44] Tongxue Zhou, Stephane Canu, Pierre Vera, and Su Ruan. Brain tumor segmentation with missing modalities via latent multi-source correlation representation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020.