Pattern Recognition 141 (2023) 109665

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Feature fusion and latent feature learning guided brain tumor segmentation and missing modality recovery network

Tongxue Zhou

School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China

ARTICLE INFO

Article history:
Received 15 August 2022
Revised 5 April 2023
Accepted 30 April 2023
Available online 2 May 2023

Keywords:
Brain tumor segmentation
Multimodal feature fusion
Missing modalities
Spatial consistency
Latent feature learning

ABSTRACT

Accurate brain tumor segmentation is an essential step for clinical diagnosis and surgical treatment. Multimodal brain tumor segmentation strongly relies on an effective fusion method and an excellent segmentation network. However, it is common to have some missing MR modalities in clinical scenarios due to image corruption, acquisition protocol, scanner availability and scanning cost, which can heavily decrease the tumor segmentation accuracy, and also cause information loss for down-streaming disease analysis. To address this issue, I propose a novel multimodal feature fusion and latent feature learning guided deep neural network. On the one hand, the proposed network can help to segment brain tumors when one or more modalities are missing. On the other hand, it can retrieve the missing modalities to compensate for incomplete data. The proposed network consists of three key components. First, a Multimodal Feature Fusion Module (MFFM) is proposed to effectively fuse the complementary information from different modalities, consisting of a Cross-Modality Fusion Module (CMFM) and a Multi-Scale Fusion Module (MSFM). Second, a Spatial Consistency-based Latent Feature Learning Module (SC-LFLM) is presented to exploit multimodal latent correlation and extract the relevant features to benefit segmentation. Third, the Multi-Task Learning (MTL) paths are integrated to supervise the segmentation and recover the missing modalities. The proposed method is evaluated on BraTS 2018 dataset, and it can achieve superior segmentation results when one or more modalities are missing, compared with the state-of-the-art methods. Furthermore, the proposed modules can be easily adapted to other multimodal network architectures and research fields.

1. Introduction

Brain tumors are the growths of abnormal cells in the brain or central spinal canal. Gliomas are the most common types of primary tumors that occur in the brain or spinal cord. Estimated 80,000 people are newly diagnosed with primary brain tumors each year in the U.S. and around 25% of these are gliomas. Gliomas can be roughly classified into two groups according to their grade: High-Grade Gliomas (HGG) and Low-Grade Gliomas (LGG). Although LGG is less aggressive than HGG, all LGG finally progress to HGG and death [1–3]. Therefore, the early diagnosis of brain tumor plays an important role in clinical practice. Diagnosing brain tumors usually begins with MRI. MRI can provide better soft tissue contrast and it is a non-invasive imaging technology. Also, we can obtain 3-dimensional (3D) images from MRI. In addition, MRI is a multimodal imaging approach. It can express various contrasts due to the various relaxation properties of the protons in the tissues. The commonly used modalities are T1-weighted (T1), contrast-enhanced T1-weighted (T1c), Fluid Attenuation Inversion Recovery (FLAIR) and T2-weighted (T2) images, shown in Fig. 1. Compared to single modalities, multi-modalities can help to extract features from different views and bring complementary information, contributing to better data representation and discriminative power of the network [4]. However, the complete MR modalities are usually unavailable in most cases due to the different acquisition protocol, image corruption, scanner availability and scanning cost.

In this paper, I propose a multimodal feature fusion and spatial consistency-based latent feature learning network to segment brain tumors as well as to recover the missing modalities. The contributions of this work can be summarized as follows:

(1) To learn useful feature representations from different modality data, a multimodal feature fusion model is presented. It consists of a cross-modality fusion module which is based on the self-attention model and a multi-scale fusion module. Through these two modules, the network can selectively emphasize informative features and suppress less useful ones.

(2) In order to reveal the intrinsic relationship between multiple modalities, I introduce a novel spatial consistency-based latent feature learning module to exploit the multimodal correlation and

E-mail address: zhoutongxue1992@163.com

https://doi.org/10.1016/j.patcog.2023.109665
0031-3203/

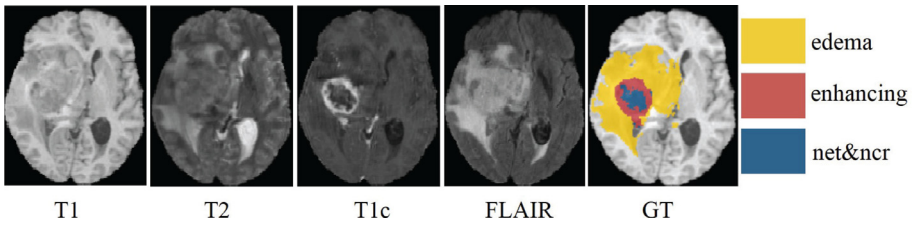

Fig. 1. The commonly used four MR modalities with ground-truth, including three sub-tumor regions: edema, enhancing tumor and non-enhancing with necrosis (net&ncr).

learn the latent correlated features. Meanwhile, the learned correlated features can be used to improve the segmentation.

(3) To achieve both brain tumor segmentation and missing data recovery in a single network, multi-task learning is introduced in this work, including a segmentation task, a reconstruction task and a generation task. The multiple learning paths can not only further supervise the target segmentation task but also generalize the overall network by sharing knowledge among different tasks.

(4) Comprehensive experiments conducted on BraTS 2018 dataset demonstrate that the effectiveness of the proposed components and the proposed method outperforms the state-of-the-art methods.

The remainder of this paper is organized as follows: Section 2 introduces the related work, Section 3 elaborates on the proposed method, and Section 4 describes the experimental setup. Section 5 presents the experimental results. Section 6 gives the conclusion of the work.

2. Related works

2.1. Brain tumor segmentation with full modalities

Brain tumor segmentation in MRI remains challenging for several reasons. For example, brain tumors can appear at variable locations with different sizes and shapes. In addition, brain tumor are very heterogeneous, and the intensity value of a brain tumor may overlap with the intensity value of the healthy brain tissue [5–7]. In recent years, deep learning has demonstrated excellent performance in a wide range of fields, such as object detection [8], visual tracking [9], regression prediction [10], image classification [11], image generation [12] and image segmentation [4]. Researchers in the medical image field have also applied deep learning to tackle brain tumor segmentation in MRI [13,14]. Based on the popular public multimodal brain tumor segmentation dataset BraTS, a large number of approaches have been proposed. For example, Kamnitsas et al. [15] proposed an ensemble of various CNNs to realize a good generalization performance, which achieves the best performance in the BraTS 2017 competition. Myronenko et al. [16] introduced the variational auto-encoder (VAE) to a U-Net-based brain tumor segmentation network. Jiang et al. [17] proposed a two-stage cascaded U-Net to refine the segmentation results gradually. Isensee et al. [18] proposed nnU-Net and incorporated some BraTS-specific modifications regarding post-processing, region-based training and data augmentation to improve the segmentation accuracy.

2.2. Brain tumor segmentation with missing modalities

Despite the recent success of brain tumor segmentation approaches, their application to some specific issues is still limited, such as segmentation in the case of missing modalities. It is difficult to always have complete modalities in clinical scenarios due to the different acquisition protocols, image corruption, scanner availability and scanning cost. In addition, the missing information can cause restraints in MRI analysis, diagnosis and research studies. Thus, recovering the missing modalities is an essential step in medical diagnosis, surgery treatment and medical research such as segmentation, detection and multimodal registration [19].

In recent years, there exists a large amount of work in the field of brain tumor segmentation with missing modalities. On the one hand, some researchers attempted to retrieve the missing information by exploiting the multimodal latent feature space. For example, Havaei et al. [20] proposed a network named HeMIS and Lau et al. [21] proposed a network named URN, both of which proposed calculation of arithmetic operations (mean and variance) to aggregate the independent features to obtain a shared latent feature representation for segmentation. Chartsias et al. [22] proposed to minimize the L1 or L2 distance between features from different modalities to achieve the latent feature representation. Dorent et al. [23] proposed a network named U-HeMIS to apply multimodal variational auto-encoders to cope with the absence of modalities. Chen et al. [24] introduced feature disentanglement to address the missing data issue. Shen et al. [25] proposed a domain adaptation approach to recover the information from the missing modality. Zhu et al. [26] proposed a cascade module to supplement the features of missing modalities. On the other hand, many works have been proposed first synthesizing the missing modalities, and then segmenting brain tumors using the existing and synthesized modalities. For example, Islam et al. [19] designed a synthesis model from multimodal MRI to single MRI modality, and achieved the segmentation using both available and synthesized modalities. However, this method can only cope with one missing modality. A similar approach can be observed in the literature [27], while the tasks of synthesis and segmentation are separately performed. In this work, the proposed approach can not only segment brain tumors with any number of missing modalities but also can retrieve the missing modalities at the same time via a single deep neural network.

2.3. Multi-task learning using deep neural networks

Multi-task learning aims to learn multiple tasks in parallel to improve generalization by sharing knowledge among tasks [28,29]. Recently, MTL has attracted much attention in deep learning communities including object detection [30,31], image classification [32] and image segmentation [33]. MTL has also extended into medical image segmentation. For example, Huang et al. [34] proposed a deep multi-task learning framework to perform distance estimation as well as tumor segmentation. Foo et al. [35] proposed to combine image classification and image segmentation for diabetic retinopathy. Amyar et al. [36] proposed a multi-task deep learning model to jointly identify COVID-19 patients and segment COVID-19 lesions from chest CT images. MTL methods can be categorized into hard and soft parameter-sharing methods. In hard parameter sharing, multiple tasks are learned with shared network layers and task-specific layers. In soft parameter sharing, each task has its own model with its own parameters, and the distance between the parameters of the model is regularized to encourage the parameters to be similar. Hard parameter sharing is the most com-

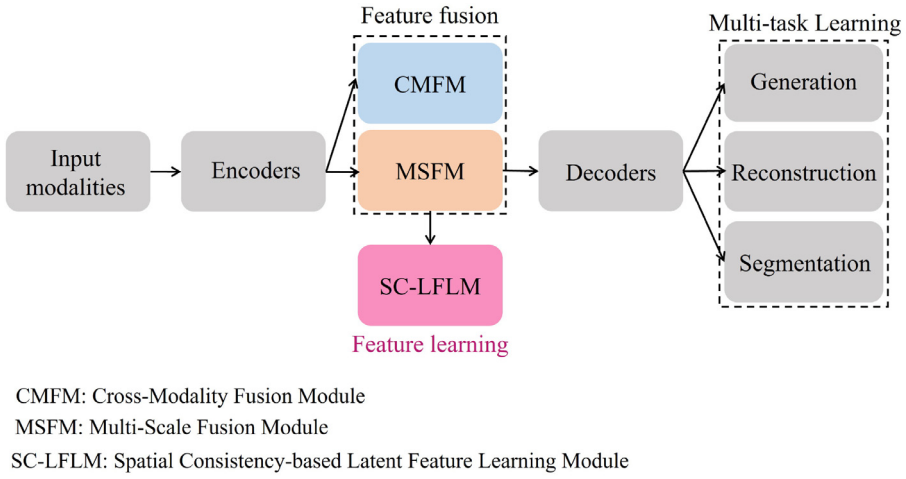

Fig. 2. The flowchart of the proposed method. Input modalities are first passed to the multi-encoders to extract independent features for each modality. Then, the feature fusion and feature learning models are applied to learn the informative and correlated features. Following that, the decoders are utilised to achieve multi-task learning.

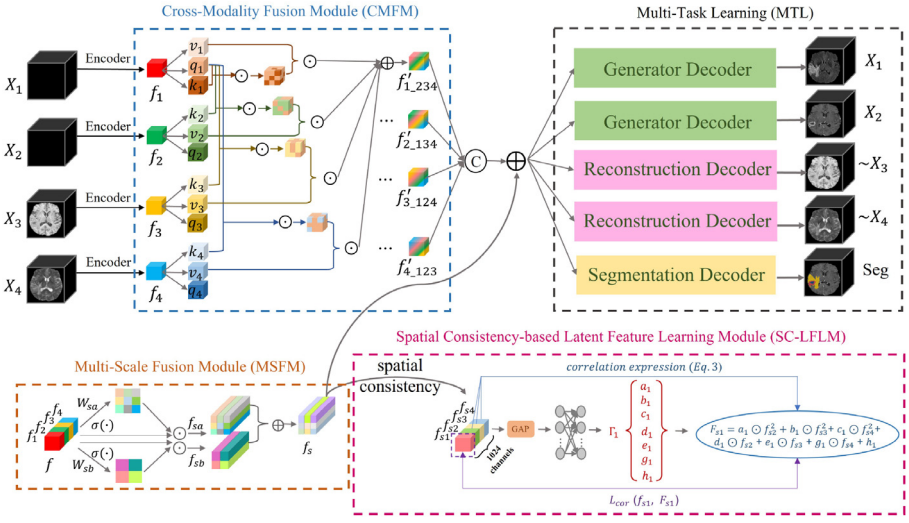

Fig. 3. The architecture of the proposed network. Here I assume X_1 and X_2 modalities are missing. The network consists of four individual encoders to extract the independent features for each modality, a cross-modality fusion model, a multi-scale fusion module, a latent feature learning module and three task-specific decoders. $\oplus$ denotes pixel-wise addition, $\odot$ denotes pixel-wise multiplication, and © denotes concatenation.

monly used approach, which can decrease the risk of over-fitting and reduce the training time compared to soft parameter sharing. In this work, the hard parameter-sharing method is employed.

3. Methodology

The flowchart of the proposed method is presented in Fig. 2. First, the available MR modalities are fed into the individual encoders to learn independent features for each modality. Then, a multimodal feature fusion model is proposed to extract the informative features for segmentation, including a cross-modality fusion module and a multi-scale fusion module. In addition, a spatial consistency-based latent feature learning module is applied to exploit the latent multimodal correlation and learn the correlated features to benefit the segmentation. Following that, the multi-task learning paths are implemented, consisting of the modality generation task for missing modalities, the modality reconstruction task for available modalities, and the brain tumor segmentation task. Multi-task learning can leverage useful information contained in the multiple related tasks to help improve the generalization performance of all the tasks [28]. The detailed network architecture is presented in Fig. 3.

3.1. Motivation

Considering that multiple modalities can provide complementary information about tumor regions from different views, I first propose a multimodal feature fusion model to selectively learn informative features. The proposed fusion model can not only learn cross-modality features but also extract multi-scale spatial contextual feature information. In addition, the same tumor regions can be observed in multiple modalities, so it is reasonable to assume that a spatial consistency exists between modalities. To capture the latent correlation between modalities, a latent feature learning module is proposed. In addition, to recover the missing modalities, multi-task learning is proposed to segment tumor regions as well as to generate missing modalities.

3.2. Multimodal feature fusion model (MFFM)

Choosing an effective feature fusion approach plays an important role in segmentation tasks [37]. For multimodal brain MR images, different MR modalities can highlight different tissue structures and underlying anatomy. For example, tumor with peritumoral edema can be obviously distinguished from T2 modality

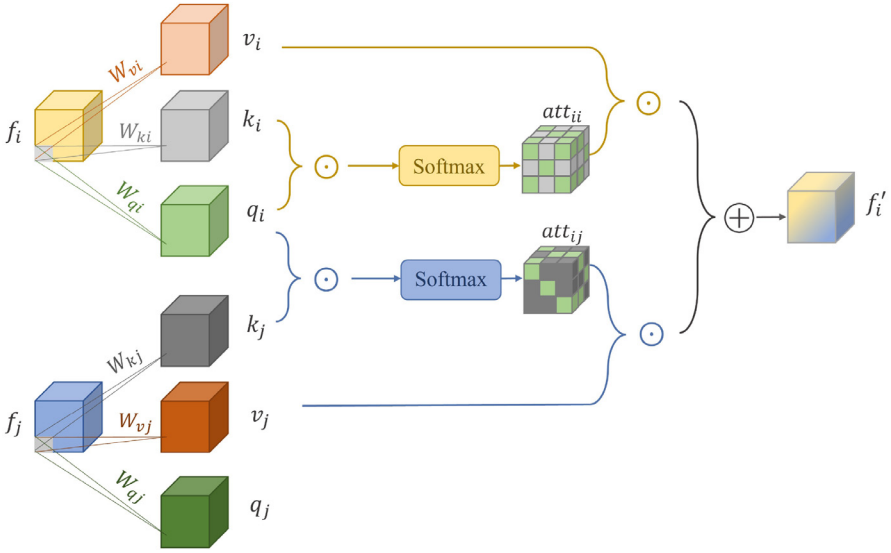

Fig. 4. The architecture of the proposed CMFM. Here I take features f_i and f_j of modality i and modality j as an example. Each modality feature is first transformed to a set of feature maps $\{q_i, k_i, v_i\}$ by three separate convolution operations. Then, the cross-modality weight att_{ij} can be computed via pixel-wise multiplication between each pair of feature maps $\{q_i, k_j\}$ followed by a softmax function. It is noted that att_{ii} is not the cross-modality attention weight, it is obtained from modality i and it can be computed via pixel-wise multiplication between feature maps q_i and k_i, followed by a softmax function. Finally, the cross-modality feature can be obtained by multiplying the cross-modality weight with the corresponding value v_i and v_j. The fused cross-modality feature $f_i^{'}$ of modality i can be obtained by a summation through all the cross-modality features.

and FLAIR modality. The enhancing tumor core region can be clearly observed in T1c modality. Therefore, a Multimodal Feature Fusion Model (MFFM) is proposed to learn the multimodal complimentary feature information, which consists of a Cross-Modality Fusion Module (CMFM) and a Multi-Scale Fusion Module (MSFM), through which the network can selectively emphasise informative features and suppress less useful ones.

3.2.1. Cross-Modality fusion module (CMFM)

The proposed Cross-Modality Fusion Module (CMFM) is inspired by the recent self-attention model [38] in machine translation. A self-attention model computes the response at a position in a sequence by learning a weighted average feature representation by considering all the positions. Convolution processes the information in a local neighbourhood which is inefficient for modelling long-range dependencies in images. To address this, the self-attention model is adopted to enable the network to learn non-local structures in the image to learn more useful feature information for segmentation. Compared with the original self-attention model, there are two improvements in the network: (1) The self-attention model is applied to exploit the cross-modal features in 3D feature representations, instead of the 1D sequences. To achieve this, the scaled dot-product attention is replaced by pixel-wise multiplication attention. (2) The multiplication operation between the input tensor and weight matrices (W_q, W_k, W_v) is replaced by the convolution operation adapting to the 3D feature representations. The architecture of the proposed CMFM is presented in Fig. 4.

First, each independent feature f_i is forwarded to three convolution blocks to obtain a series of feature maps $\{q_i, k_i, v_i\}$, $q_i = W_{qi} * f_i$, $k_i = W_{ki} * f_i$, $v_i = W_{vi} * f_i$, where W_{qi}, W_{ki} and W_{vi} are the convolution weights, $*$ is the convolution operation. The cross-modality attention weight att_{ij} between modality i and modality j can be computed via pixel-wise multiplication between q_i and k_j, followed by a softmax function to normalize the weight.

$$att_{ij} = softmax(q_i \odot k_j) \tag{1}$$

where q_i, k_j are the feature maps of modality i and modality j, respectively. $\odot$ is the element-wise multiplication, and att_{ij} is the cross-modality attention weight between modality i and modality j.

Then, the cross-modality feature can be calculated between cross-modality attention weight att_{ij} and the other corresponding value v_j via pixel-wise multiplication. Finally, the fused cross-modality feature $f_i^{'}$ can be obtained by summing up all the cross-modality features. In this way, the network can learn cross-modality features, which can enhance the important features and also suppress the weak ones.

$$f_i^{'} = \sum_{i,j=1}^{n} att_{ij} v_j \tag{2}$$

where $f_i^{'}$ is the fused cross-modality feature, att_{ij} is the cross-modality attention weight between modality i and modality j, it is noted that att_{ii} is not the cross-modality attention weight, it is obtained from modality i. v_j is the learned feature map from modality j.

3.2.2. Multi-Scale fusion module (MSFM)

The spatial feature information is particularly important for the segmentation task. In addition, different scale features can provide different receptive fields for the network, which can capture more crucial information for segmentation. To achieve this, I propose a Multi-Scale Fusion Module (MSFM) to explore the multimodal spatial feature information. The architecture of the proposed MSFM is depicted in Fig. 5. In the proposed MSFM, the independent features (f_1, f_2, f_3, f_4, ..., f_n) are first concatenated as $f = [f_1, f_2, f_3, f_4, \ldots, f_n]$, in this work, n=4. Then, two convolution operations with different kernel sizes ($1 \times 1 \times 1$ and $3 \times 3 \times 3$) are used to capture the multi-scale feature information for all the modalities: $s_a = W_{sa} * f$, $s_b = W_{sb} * f$, where W_{sa} and W_{sb} are the convolution weights, and they are modality specific. Here, the choice of convolution kernel size is based on the brain tumor size, the diameter of brain tumor is usually around 2.5 cm, and the small kernel size can capture the pixel-wise features to help segmentation. Following that, a sigmoid function is used to obtain the space-wise weights $\sigma(s_a)$ and $\sigma(s_b)$. The two space-wise weights $\sigma(s_a)$ and $\sigma(s_b)$ are then multiplied with the input feature f

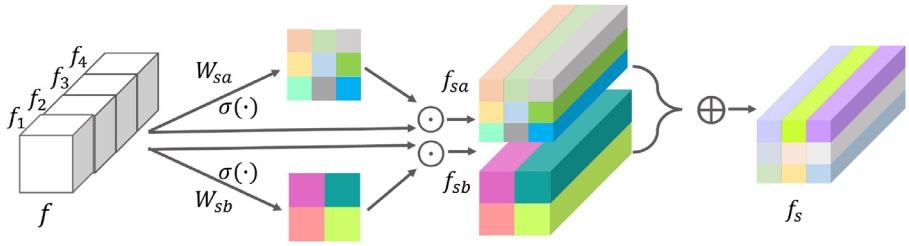

Fig. 5. The architecture of the proposed MSFM, the concatenated feature f is first passed to two convolution layers with a sigmoid function separately to learn the space-wise weights. Then, these weights are multiplied with the input feature to obtain the features from different receptive fields f_{sa}, f_{sb}. Finally, the two features are added together to achieve the fused multi-scale feature f_s.

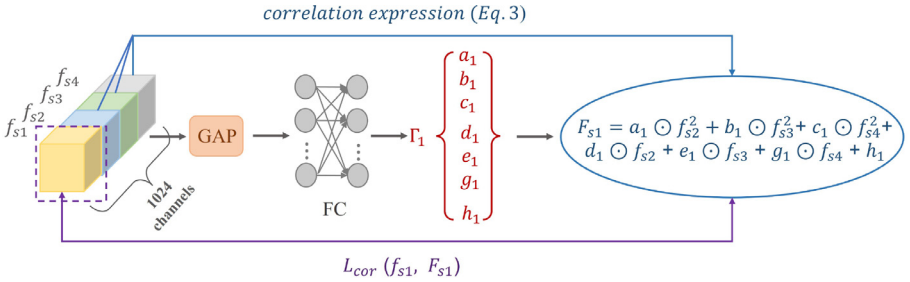

Fig. 6. The architecture of the proposed SC-LFLM. The concatenated features are first fed to a global average pooling to learn the overall feature. Then, they are passed through two fully connected layers with LeakyReLU. A set of correlation parameters can be obtained. Following that, the correlated feature can be obtained via correlation expression. In addition, correlation constraint loss is proposed to ensure the two distributions of the features is as close as possible.

to achieve the multi-scale features: $f_{sa} = \sigma(s_a)f$ and $f_{sb} = \sigma(s_b)f$. The final fused feature is obtained by summing the two multi-scale features $f_s = f_{sa} + f_{sb}$. With the assistance of the proposed MSFM, the network can learn multi-scale spatial contextual feature information between modalities to further improve the segmentation.

3.3. Spatial consistency-based latent feature learning module (SC-LFLM)

For the same patient, we can obtain different MR modalities, and these MR modalities can present different characteristics for the same tumor regions. It is reasonable to assume that there exists a spatial consistency between modalities, indicating there is a correlation on the same tumor regions between different modalities. By investigating the joint intensities of the MR images [39], we can observe a nonlinear correlation in intensity distribution between each pair of modalities. To exploit this correlation, a Spatial Consistency-based Latent Feature Learning Module (SC-LFLM) is proposed. The architecture of the proposed SC-LFLM is illustrated in Fig. 6. Through the MSFM module, a multi-scale fused feature f_s can be obtained. It is noted that f_s consists of four features: f_{s1}, f_{s2}, f_{s3}, f_{s4}, and each one represents a multi-scale feature, e.g. f_{s1} represents the multi-scale feature of f_1. First, a global average pooling (GAP) followed by two fully connected layers is used to map the multi-scale fused feature f_s (1024 channels) to a set of correlation parameters for each of f_{s1}, f_{s2}, f_{s3}, f_{s4}. Here I take f_{s1} as an example, and the correlation parameters are $\Gamma_1 = \{a_1, b_1, c_1, d_1, e_1, g_1, h_1\}$. These correlation parameters describe the relationships between multi-modalities. Then, the correlated feature F_{s1} of modality X_1 can be achieved via a nonlinear correlation expression (Equation 3). Finally, the Kullback–Leibler divergence-based correlation constraint loss (Equation 4) is proposed to measure the similarity between the estimated correlated feature and the original feature of modality X_1, the lower loss will attribute to higher multimodal similarity. It is noticed that the abovementioned CMFM module is proposed to learn cross-modality features. MSFM module is proposed to learn multi-scale spatial contextual features. SC-LFLM module is based on the MSFM module, however, it is proposed to exploit the correlation among different modalities based on the same tumor region. In addition, a correlation loss function is employed to encourage the network to learn latent correlated features to benefit segmentation.

$$F_{s1} = a_1 \odot f_{s2}^2 + b_1 \odot f_{s3}^2 + c_1 \odot f_{s4}^2 + d_1 \odot f_{s2} + e_1 \odot f_{s3} + g_1 \odot f_{s4} + h_1 \tag{3}$$

where F_{s1} is the correlated feature, and f_{s2}, f_{s3}, f_{s4} are the original features, and a_1, b_1, c_1, d_1, e_1, g_1, h_1 are the correlation parameters.

$$L_{cor_1} = P(f_{s1}) \log \frac{P(f_{s1})}{Q(F_{s1})} \tag{4}$$

where $P(f_{s1})$ and $Q(F_{s1})$ are probability distributions of the original feature and the correlated feature from modality X_1, and they are estimated during training.

The total correlation loss function is defined as:

$$L_{cor} = L_{cor_1} + L_{cor_2} + L_{cor_3} + L_{cor_4} \tag{5}$$

where L_{cor_1}, L_{cor_2}, L_{cor_3} and L_{cor_4} are the correlation loss functions for modality X_1, X_2, X_3 and X_4, respectively.

3.4. Multi-Task learning (MTL)

The proposed multi-task learning consists of three tasks: image reconstruction for available modalities, image generation for missing modalities, and image segmentation. The three tasks share the same encoders and own their task-specific decoders, which allows the individual tasks to learn a shared feature representation, as well as to improve the generalization of the network. In addition, the auxiliary tasks (image reconstruction and generation) can improve the performance of the target task (image segmentation). The architecture of the network is depicted in Fig. 7. Specifically, in the encoder, each layer includes a $3 \times 3 \times 3$ convolution with $stride = 2$ and a Res_dil block [40] except the first layer (the grey block) where $stride = 1$. In the decoder, each layer consists of a 3D upsampling layer, a $3 \times 3 \times 3$ convolution and a Res_dil block. In

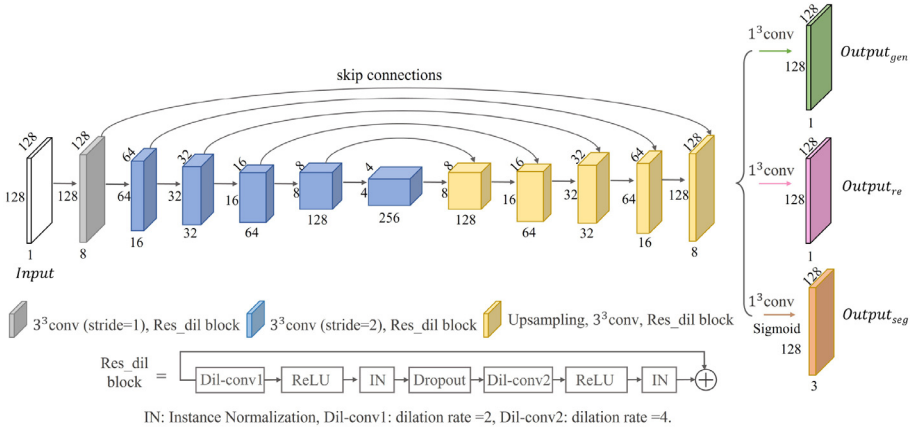

Fig. 7. The architecture of the encoder-decoder for multi-task learning, where $Output_{gen}$, $Output_{re}$ and $Output_{seg}$ denote the outputs of generation, reconstruction and segmentation tasks, respectively.

addition, the layers in the encoder are skip-connected and concatenated with the corresponding layers in the decoder, which allows the network to recover spatial information caused by downsampling and obtain the fine-grained details.

3.5. Loss function

The training loss function is defined in Equation 6, which consists of three terms: L_{seg}, L_1 and L_{cor}. L_{seg} is the segmentation network, L_1 is the generation and reconstruction loss, and L_{cor} is the latent feature learning module loss, which is presented in Equation 5.

$$L_{total} = L_{seg} + \xi L_1 + \psi L_{cor} \tag{6}$$

where ξ and ψ are the trade-off parameters, which are set empirically as 0.1.

The segmentation loss function is based on the Dice loss, which measures the overlap between the prediction region and the ground truth region.

$$L_{seg} = 1 - 2\frac{\sum_{i=1}^{C}\sum_{j=1}^{N} p_{ij}g_{ij} + \epsilon}{\sum_{i=1}^{C}\sum_{j=1}^{N}(p_{ij} + g_{ij}) + \epsilon} \tag{7}$$

where N indicates the number of pixels in the image, C is the number of the classes, $p_{ij} \in [0, 1]$ is the output probability of pixel i for class j, $g_{ij} \in \{0, 1\}$ is the ground truth labelling of pixel i for class j, and ϵ is a small constant to avoid dividing by 0.

The generation and reconstruction loss functions are based on the L_1 loss, which compares the difference between the predicted image and the ground-truth image.

$$L_1 = \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{8}$$

where N is the number of pixels in the image, y is the ground-truth image, and $\hat{y}$ is the generated image.

4. Experimental setup

4.1. Dataset and implementation details

The proposed method is evaluated on the public multimodal brain tumor segmentation dataset BraTS 2018 [41], which contains 285 cases with ground-truth, each case has four MR modalities including T1, FLAIR, T1c and T2. There are three segmentation labels: Whole Tumor (WT), Tumor Core (TC) and Enhancing Tumor (ET). The whole Tumor consists of all tumor tissues, Tumor Core consists of enhancing tumor, necrotic and non-enhancing tumor core. All the provided data have been pre-processed by organisers, including co-registering to the same anatomical template, interpolating to the same resolution ($1mm^3$) and skull-stripping. The ground-truth labels have been manually labelled by experts. In this work, the images are resized from $155 \times 240 \times 240$ to $128 \times 128 \times 128$. Bias field correction is corrected by using the N4ITK tool. Each image is normalized to a zero-mean, unit-variance space.

Table 1
The parameter settings of the network.

Parameters	Value
Layer	6
Input size	$128 \times 128 \times 128$
Initial filter	8
Initial learning rate	0.0005
Optimizer	Ndame
Batch size	1
Training samples	285
Segmentation labels (C)	3
ψ	0.1
ξ	0.1

The proposed network is implemented with Keras using a single Nvidia Tesla V100 (32G). Nadam is used as the optimizer, the initial learning rate is set as 0.0005, which will be halved after 5 epochs if the validation loss is not improved. Early stopping is used to avoid over-fitting, where the training will stop if the validation loss is not improved over 10 epochs. The dataset is randomly split into 80% training and 20% testing. The experimental results are obtained by submitting the local results to the online evaluation platform[1]. More details about the parameters are described in Table 1.

4.2. Evaluation metrics

4.2.1. Segmentation evaluation metrics

Two evaluation metrics are applied to calculate the segmentation performance, including Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD), and a higher value of DSC and a lower value of HD are considered as the better results.

$$DSC = \frac{2|V_p \cap V_g|}{|V_p| + |V_g|} \tag{9}$$

where V_p and V_g denote the set of prediction and ground truth pixels for a given class, and $|\cdot|$ denote the volume of the enclosed

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

Table 2
Comparison results in terms of DSC between different methods on the BraTS 2018 dataset. Higher DSC values indicate better results. • denotes the included modality and ◦ denotes the missing one, bold results denote the best scores. WT, TC, and ET denote whole tumor, tumor core and enhancing tumor, respectively. AVG denotes the average results on the three target regions, Mean denotes the average results on one target region across all the situations. $*$ denotes the significant improvement evaluated via the Wilcoxon test ($p < .05$).

Modality				Baseline				+ CMFM				+ CMFM + MSFM				+ CMFM + MSFM + SC-LFLM			
F	T1	T1c	T2	WT	TC	ET	AVG	WT	TC	ET	AVG	WT	TC	ET	AVG	WT	TC	ET	AVG
◦	◦	◦	•	75.1	45.7	28.2	49.7	71.5	45.7	16.1	44.4	70.9	40.0	17.1	42.7	**80.3***	**55.6***	**33.9***	**56.6**
◦	◦	•	◦	65.2	77.2	71.8	71.4	68.0*	80.4*	74.2*	74.2	68.8	81.9	**75.1**	75.2	**70.3**	**82.9***	74.9	**76.0**
◦	•	◦	◦	63.0	39.4	17.9	40.1	64.9*	45.0*	12.8	40.9	69.3*	47.7	17.7	44.9	**73.7***	**57.6***	**33.6***	**55.0**
•	◦	◦	◦	82.2	52.3	28.3	54.3	82.2	55.3*	24.3	53.9	83.3	56.5	27.5	55.8	**84.9***	**63.0***	**38.5***	**62.1**
◦	◦	•	•	80.2	79.8	74.3	78.1	80.2	84.5*	77.1*	80.6	78.3	84.4	77.8	80.2	**81.7***	**86.7***	**78.2***	**82.2**
◦	•	•	◦	71.9	78.2	74.6	74.9	72.8	84.0*	76.0*	77.6	73.9*	84.1	**77.4**	78.5	**75.5**	**85.1***	76.8*	**79.1**
•	•	◦	◦	83.5	55.5	33.0	57.3	84.0	59.6*	32.3	58.6	85.2	61.0	32.7	59.6	**86.4***	**66.3***	**42.5***	**65.1**
◦	•	◦	•	78.7	49.1	31.2	53.0	77.4	53.2*	21.9	50.8	77.7*	55.0*	24.8*	52.5	**82.3***	**62.1***	**39.4***	**61.3**
•	◦	◦	•	83.5	53.1	34.1	56.9	83.3	56.6*	31.6	57.2	85.1*	59.3	34.2	59.5	**85.9**	**65.1***	**43.7***	**64.9**
•	◦	•	◦	82.6	80.6	76.2	79.8	85.2*	84.7*	77.6*	82.5	84.7	84.3	78.4	82.5	**85.8***	**86.5***	**78.9***	**83.7**
•	•	•	◦	83.6	81.5	76.7	80.6	85.2*	85.4*	78.0*	82.9	85.0	85.1	78.7	82.9	**86.6***	**87.1***	**78.8**	**84.1**
•	•	◦	•	84.1	55.3	35.4	58.3	84.2	58.1*	34.6	59.0	85.6*	62.3*	37.6*	61.8	**86.5**	**67.0***	**45.3***	**66.3**
•	◦	•	•	84.2	81.8	76.2	80.8	85.5*	84.7*	77.6*	82.6	85.3	84.7	78.5	82.9	**86.4***	**86.7***	**78.8***	**83.9**
◦	•	•	•	81.3	80.5	75.4	79.1	80.0	85.0*	77.5*	80.8	79.7	85.1	**78.2**	81.0	**82.6***	**86.8***	78.1	**82.5**
•	•	•	•	84.4	82.2	76.6	81.1	85.4*	85.2*	77.8*	82.8	85.1	85.1	**78.6**	82.9	**86.5***	**87.0***	**78.6**	**84.1**
Mean				78.9	66.1	54.0	66.4	79.3	69.8	52.6	67.3	79.9	70.4	54.3	68.2	**82.4**	**75.0**	**60.0**	**72.5**

set.

$$HD = \max\{\max_{s\in S}\min_{r\in R} d(s,r), \max_{r\in R}\min_{s\in S} d(r,s)\} \tag{10}$$

where S and R are the two sets of the surface points of the prediction and the real annotation, respectively, and d is the Euclidean distance.

4.2.2. Generation evaluation metrics

Three evaluation metrics are applied to calculate the generation performance, including Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). In addition, Wilcoxon signed-rank test is applied to see the importance of the proposed components. If the p-value is lower than 0.05, it means there are significant improvements by using the proposed components, which is denoted by $*$ in the tables.

MSE is the simplest and most widely used quality metric, which measures the average of the square of the errors between generated image and real image. It is calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{11}$$

where n is the number of pixels in the image, y and $\hat{y}$ are the ground-truth image and the generated image, respectively.

PSNR is applied to measure the prediction accuracy in terms of the logarithmic decibel scale. The larger PSNR indicates that the generation is of higher quality. It is defined as:

$$PSNR = 10\log_{10}\frac{Max}{MSE} \tag{12}$$

where Max is the maximum pixel value in the image.

SSIM is a perception-based model that considers image degradation as perceived change in structural information. The larger SSIM indicates that the generation is of higher quality. It is computed as:

$$SSIM = \frac{(2\mu_{\hat{y}}\mu_y + c_1)(2\sigma_{y\hat{y}} + c_2)}{(\mu_{\hat{y}}{}^2 + \mu_y^2 + c_1)(\sigma_{\hat{y}}{}^2 + \sigma_y^2 + c_2)} \tag{13}$$

where μ_y and σ_y^2 are the mean and variance of the ground-truth image y, $\mu_{\hat{y}}$ and $\sigma_{\hat{y}}^2$ are the mean and variance of the generated image $\hat{y}$, and $\sigma_{y\hat{y}}$ is the covariance between the ground-truth image y and the generated image $\hat{y}$. c_1 and c_2 are to stabilize the division with weak denominator.

5. Experiment results and analysis

5.1. Ablation experiments

To analyze the effectiveness of the proposed strategies, the ablation experiments are conducted based on a baseline method. The baseline is the proposed method without using CMFM, MSFM and SC-LFLM. From Table 2, first, it can be observed that an additional input modality can result in statistically significant improvements. In addition, the proposed CMFM can improve the baseline by 1.4% in terms of average DSC, especially when only T1c modality is available, a 3.9% improvement in terms of average DSC can be observed. It can be explained that the CMFM can capture the cross-modality feature information to help tumor segmentation. By comparing "Baseline" and "Baseline + CMFM + MSFM", it can be observed that the segmentation accuracy is further boosted with a 2.7% improvement in terms of average DSC. Because the MSFM can further improve the feature learning ability of the model by considering the multi-scale feature information. When the SC-LFLM is integrated, the method can obtain a 9.2% performance gain in terms of average DSC compared with the baseline method. A significant increase of 37.2% in terms of average DSC can be observed compared with the baseline method when only the T1 modality exists. It demonstrates that the proposed latent feature learning module can extract important feature information to aid brain tumor segmentation. The Wilcoxon test results also show that the proposed method can achieve significantly better accuracy than other compared methods.

Besides, I also study the segmentation accuracy in terms of HD in Table 3, from which it can be observed that the proposed CMFM, MSFM and SC-LFLM can obtain improvements of 4.7%, 16.8%, and 34.6% in terms of average HD compared with the baseline, respectively. More encouragingly, the significant improvements (denoted by $*$) can be observed via the Wilcoxon test, which further reveals the advantage of the proposed strategies.

5.2. Comparison with single-Task learning (STL) and dual-Task learning (DTL)

Then, to validate the effectiveness of the proposed MTL (Multi-Task Learning), I compare it with STL (Single-Task Learning), where

Table 3
Comparison results in terms of HD between different methods on BraTS 2018 dataset. Lower HD values indicate better results. • denotes the included modality and ○ denotes the missing one, bold results denote the best scores. WT, TC, and ET denote whole tumor, tumor core and enhancing tumor, respectively. AVG denotes the average results on the three target regions, Mean denotes the average results on one target region across all the situations. ∗ denotes the significant improvement evaluated via the Wilcoxon test ($p < .05$).

Modality				Baseline				Baseline + CMFM				Baseline + CMFM + MSFM				Baseline + CMFM + MSFM + SC-LFLM			
F	T1	T1c	T2	WT	TC	ET	AVG	WT	TC	ET	AVG	WT	TC	ET	AVG	WT	TC	ET	AVG
○	○	○	•	13.9	19.9	18.7	17.5	22.5	24.2	22.2	23.0	14.0*	18.9	18.0	17.0	**7.7***	**12.7***	**11.6***	**10.7**
○	○	•	○	13.9	9.3	7.1	10.1	13.7	11.2	7.4	10.7	13.1*	7.2	5.6	8.6	**10.1**	**6.8**	**5.5**	**7.5**
○	•	○	○	19.1	26.0	23.0	22.7	16.7	17.4	18.4	17.5	13.1*	17.6	16.3*	15.6	**10.0***	**15.4**	**14.1**	**13.2**
•	○	○	○	9.1	15.9	13.6	12.9	9.7	16.1	14.1	13.3	9.5*	13.9	12.6	12.0	**5.2**	**11.8**	**10.9**	**9.3**
○	○	•	•	8.9	8.0	6.2	7.7	7.9	4.8	3.4	5.4	8.9	5.1	3.8	6.0	**6.5***	**4.0***	**2.8***	**4.4**
○	•	•	○	10.9	8.9	5.7	8.5	11.3	6.0	**3.9**	7.1	9.6	**5.5**	4.3	6.4	**8.4**	5.6	4.3	**6.1**
•	•	○	○	7.4	13.7	12.5	11.2	8.2	13.5	12.2	11.3	7.6	12.9*	11.5*	10.6	**5.8**	**10.8**	**10.9**	**9.2**
○	•	○	•	10.6	18.7	16.6	15.3	13.5	15.6	14.3	14.5	10.0	15.2	14.1	13.1	**6.4***	**11.6**	**11.3**	**9.8**
•	○	○	•	6.7	16.0	15.0	12.5	9.4	16.1	14.6	13.4	7.4	11.6*	10.0*	9.7	**4.9**	**11.4**	**9.9**	**8.7**
•	○	•	○	7.3	5.4	3.9	5.5	6.5	4.8	3.4	4.9	6.9	4.4	3.0	4.8	**5.6***	**3.9***	**2.5**	**4.0**
•	•	•	○	6.4	5.5	3.8	5.2	6.7	4.5	3.0	4.7	6.2*	4.6	3.1	4.7	**5.3***	**3.5***	**2.4**	**3.7**
•	•	○	•	6.7	15.3	14.2	12.0	7.7	13.4	12.5	11.2	6.8	12.1*	10.7*	9.9	**4.6**	**9.6**	**9.2**	**7.8**
•	○	•	•	6.3	6.9	5.4	6.2	5.8	4.9	3.2	4.6	6.1	4.5*	3.0*	4.5	**4.5***	**3.6***	**2.5***	**3.5**
○	•	•	•	8.5	8.0	5.8	7.4	9.3	5.9*	4.6*	6.6	8.4	4.8	3.4	5.5	**6.2***	**4.1***	**2.8***	**4.4**
•	•	•	•	6.6	7.0	5.7	6.4	5.8	4.7	3.0	4.5	6.2	4.7	3.1	4.7	**4.6***	**3.5***	**2.3***	**3.5**
Mean				9.5	12.3	10.5	10.7	10.3	10.9	9.3	10.2	8.9	9.5	8.2	8.9	**6.4**	**7.9**	**6.9**	**7.0**

Table 4
Comparison results among STL, DTL and MTL in terms of DSC on BraTS 2018 dataset. • denotes the included modality and ○ denotes the missing one, bold results denote the best scores. WT, TC, and ET denote whole tumor, tumor core and enhancing tumor, respectively. AVG denotes the average results on the three target regions, Mean denotes the average results on one target region across all the situations. ∗ denotes the significant improvement evaluated via the Wilcoxon test ($p < .05$).

Modality				STL				DTL (reconstruction)				DTL (generation)				MTL			
F	T1	T1c	T2	WT	TC	ET	AVG	WT	TC	ET	AVG	WT	TC	ET	AVG	WT	TC	ET	AVG
○	○	○	•	78.2	51.1	20.6	50.0	79.1	51.1	31.0	53.7	79.2	53.7	31.3	54.7	**80.3**	**55.6***	**33.9***	**56.6**
○	○	•	○	**70.7**	**84.1**	**76.8**	**77.2**	69.1	82.6	74.3	75.3	69.3	82.9	74.9	75.7	70.3	82.9	74.9	76.0
○	•	○	○	69.7	49.9	17.3	45.6	72.0	54.1*	30.9*	52.3	72.1	54.0*	30.2*	52.1	**73.7***	**57.6***	**33.6***	**55.0**
•	○	○	○	84.6	**63.1**	27.5	58.4	84.0	60.3	34.9	59.7	84.0	61.0	35.8	60.3	**84.9**	63.0	**38.5**	**62.1**
○	○	•	•	80.1	85.4	77.6	81.0	80.7	**86.7***	78.3*	81.9	81.5*	86.5*	**78.4***	82.1	**81.7***	**86.7***	78.2*	**82.2**
○	•	•	○	74.1	**85.7**	**77.7**	79.2	74.2	85.4	77.2	78.9	75.4	85.5	77.2	**79.4**	**75.5**	85.1	76.8	79.1
•	•	○	○	85.9	**67.1**	35.7	62.9	86.2	64.6	40.2	63.7	**86.4**	64.3	40.2	63.6	**86.4**	66.3	**42.5**	**65.1**
○	•	○	•	81.3	59.8	29.8	57.0	81.6	59.2	37.6	59.5	81.9	58.8	35.8	58.8	**82.3**	**62.1**	**39.4**	**61.3**
•	○	○	•	**86.0**	65.0	34.9	62.0	85.3	62.6	41.5	63.1	85.4	63.7	41.6	63.6	85.9	**65.1**	**43.7**	**64.9**
•	○	•	○	85.1	86.3	78.3	83.2	85.7	85.9	78.7	83.4	**86.0***	86.0	78.7	83.6	85.8*	**86.5**	**78.9**	**83.7**
•	•	•	○	85.3	86.4	77.9	83.2	86.4	86.9	**78.8**	84.0	**86.7***	86.7	78.7	84.0	86.6*	**87.1**	**78.8**	**84.1**
•	•	○	•	86.3	**67.3**	39.7	64.4	86.3	65.3	43.5	65.0	86.3	65.5	42.9	64.9	**86.5**	67.0	**45.3**	**66.3**
•	○	•	•	85.9	86.0	77.8	83.2	86.1	86.5	78.6	83.7	**86.4***	86.1	78.6	83.7	**86.4***	**86.7**	**78.8**	**83.9**
○	•	•	•	81.0	85.9	77.4	81.5	81.7	**86.9**	**78.2**	82.3	82.4*	86.8	78.1	82.4	**82.6***	86.8	78.1	**82.5**
•	•	•	•	85.9	86.0	77.6	83.2	86.3*	86.9	**78.7**	84.0	**86.6***	86.8	78.6	84.0	86.5*	**87.0**	78.6	**84.1**
Mean				81.3	73.9	55.1	70.1	81.6	73.7	58.8	71.4	82.0	73.9	58.7	71.5	**82.4**	**75.0**	**60.0**	**72.5**

only the target segmentation task is applied, and DTL (Dual-Task Learning), where both target segmentation task and reconstruction/generation task are applied. The comparison results are presented in Table 4 and Table 5. It can be observed that, with the assistance of the reconstruction task, the segmentation result of STL is improved by 1.9% in terms of average DSC. In addition, the generation task can improve STL by 2.0% in terms of average DSC. However, there is a slight decrease in terms of average HD on both DTL tasks. Finally, when both reconstruction and generation tasks are applied to the segmentation task, an improvement of 3.4% in terms of average DSC and 6.7% in terms of average HD can be observed. It can be explained that the multiple tasks can help to learn more valuable feature information, and also provide some supervision to the target task, leading to better segmentation results. Therefore, the comparison results in Table 4 and Table 5 demonstrate the effectiveness of integrating additional tasks into the target task. However, regarding to the computational cost, the STL use 100M around trainable parameters, and the MTL use 142M around trainable parameters. The two auxiliary tasks take 42% more training parameters than STL. In future work, I will consider improving the network architecture to decrease trainable parameters and reduce computational costs.

5.3. Comparison with the state-of-the-art segmentation methods

I also compare the proposed method with several state-of-the-art methods, which have been introduced in Section 2 as well as with the U-HVED method from [23]. The comparison results are reported in Table 6, and the results on HeMIS and U-HeMIS are cited from the work [23]. From Table 6, it can be observed that the U-Net-based method (U-HeMIS) can achieve better results than the CNN-based network (HeMIS). Second, fusing multi-modalities by Variational Auto-Encoder (VAE) (U-HVED) can further improve the segmentation accuracy than simply calculating the mean and variance from the independent features (U-HeMIS). In addition, compared with the current best method [23] (U-HVED), the proposed method can obtain 12.1% improvement in terms of average DSC, which indicates the proposed spatial consistency-based latent feature learning can further learn the informative features than VAE, and it attributes to better segmentation results.

5.4. Evaluation of the generation results

Finally, in order to demonstrate that the proposed method can provide good generation results, I evaluate the generation

Table 5
Comparison results among STL, DTL and MTL in terms of HD on BraTS 2018 dataset. • denotes the included modality and ◦ denotes the missing one, bold results denote the best scores. WT, TC, and ET denote whole tumor, tumor core and enhancing tumor, respectively. AVG denotes the average results on the three target regions, Mean denotes the average results on one target region across all the situations. ∗ denotes the significant improvement evaluated via the Wilcoxon test ($p < .05$).

Modality				STL				DTL (reconstruction)				DTL (generation)				MTL			
F	T1	T1c	T2	WT	TC	ET	AVG	WT	TC	ET	AVG	WT	TC	ET	AVG	WT	TC	ET	AVG
◦	◦	◦	•	9.7	13.8	13.6	12.4	8.9	13.7	12.7	11.8	9.2	13.3	12.2	11.6	**7.7**∗	**12.7**	**11.6**	**10.7**
◦	◦	•	◦	**10.0**	**5.2**	**4.0**	**6.4**	12.1	9.0	7.7	9.6	12.4	8.0	6.7	9.0	10.1	6.8	5.5	7.5
◦	•	◦	◦	14.8	17.0	16.1	16.0	10.9∗	**14.8**	**13.5**	**13.1**	10.9∗	15.1	13.8	13.3	**10.0**∗	15.4	14.1	13.2
•	◦	◦	◦	6.8	**10.7**	**10.3**	**9.3**	6.9	15.2	13.7	11.9	6.8	14.5	13.3	11.5	**5.2**∗	11.8	10.9	**9.3**
◦	◦	•	•	7.0	4.5	3.2	4.9	6.7	4.4	3.0	4.7	6.7	4.4	3.2	4.8	**6.5**∗	**4.0**	**2.8**	**4.4**
◦	•	•	◦	9.2	**4.8**	**3.6**	5.9	**8.4**	5.1	**3.6**	**5.7**	**8.4**	5.2	3.8	5.8	**8.4**	5.6	4.3	6.1
•	•	◦	◦	6.2	**9.7**	**8.8**	**8.2**	6.0	11.0	11.1	9.4	5.9	11.7	11.2	9.6	**5.8**	10.8	10.9	9.2
◦	•	◦	•	7.3	12.0	**10.8**	10.0	8.5	12.5	10.9	10.6	8.0	12.3	12.2	10.8	**6.4**∗	**11.6**	11.3	**9.8**
•	◦	◦	•	6.2	**9.9**	**9.0**	**8.4**	6.3	13.5	12.2	10.7	6.1	13.2	10.9	10.1	**4.9**	11.4	9.9	8.7
•	◦	•	◦	7.2	4.0	2.9	4.7	6.0	4.5	2.7	4.4	5.7∗	**3.8**	**2.5**	**4.0**	**5.6**∗	3.9	**2.5**	**4.0**
•	•	•	◦	6.3	4.0	2.9	4.4	4.7∗	3.6	2.5	**3.6**	**4.6**∗	3.7	**2.4**	**3.6**	5.3∗	**3.5**	**2.4**	3.7
•	•	◦	•	6.3	**9.1**	**8.4**	7.9	5.0	10.6	10.5	8.7	5.0	10.1	10.7	8.6	**4.6**	9.6	9.2	**7.8**
•	◦	•	•	5.4	4.0	2.9	4.1	4.8∗	4.0	2.9	3.9	4.7∗	4.1	2.8	3.9	**4.5**∗	**3.6**	**2.5**	**3.5**
◦	•	•	•	8.0	4.5	3.3	5.3	7.0∗	4.7	**2.8**	4.8	6.6	4.2	2.9	4.6	**6.2**∗	**4.1**	**2.8**	**4.4**
•	•	•	•	5.5	4.1	2.9	4.2	4.7∗	3.9	2.4	3.7	4.7∗	4.3	2.8	3.9	**4.6**∗	**3.5**	**2.3**	**3.5**
Mean				7.7	**7.8**	**6.8**	7.5	7.1	8.7	7.5	7.8	7.0	8.5	7.4	7.7	**6.4**	7.9	6.9	**7.0**

Table 6
Comparison of different methods in terms of DSC on BraTS 2018 dataset. • denotes the included modality and ◦ denotes the missing one, bold results denote the best scores. WT, TC, and ET denote whole tumor, tumor core and enhancing tumor, respectively. AVG denotes the average results on the three target regions, Mean denotes the average results on one target region across all the situations.

Modality				HeMIS [20]				U-HeMIS [23]				URN [21]				U-HVED [23]				Ours			
F	T1	T1c	T2	WT	TC	ET	AVG	WT	TC	ET	AVG	WT	TC	ET	AVG	WT	TC	ET	AVG	WT	TC	ET	AVG
◦	◦	◦	•	38.6	19.5	0.0	19.4	79.2	50.0	23.3	50.8	77.5	43.6	20.3	47.1	**80.9**	54.1	30.8	55.3	80.3	**55.6**	**33.9**	**56.6**
◦	◦	•	◦	2.6	6.5	11.1	6.7	58.5	58.5	60.8	59.3	62.2	58.5	55.8	58.8	62.4	66.7	65.5	64.9	**70.3**	**82.9**	**74.9**	**76.0**
◦	•	◦	◦	0.0	0.0	0.0	0.0	54.3	37.9	12.4	34.9	50.4	34.2	19.1	34.6	52.4	37.2	13.7	34.4	**73.7**	**57.6**	**33.6**	**55.0**
•	◦	◦	◦	55.2	16.2	6.6	26.0	79.9	49.8	24.9	51.5	84.8	50.4	23.6	52.9	82.1	50.4	24.8	52.4	**84.9**	**63.0**	**38.5**	**62.1**
◦	◦	•	•	48.2	45.8	55.8	49.9	81.0	69.1	68.6	72.9	80.3	68.9	67.6	72.3	**82.7**	73.7	70.2	75.5	81.7	**86.7**	**78.2**	**82.2**
◦	•	•	◦	15.4	30.4	42.6	29.5	63.8	64.0	65.3	64.4	69.8	65.9	66.5	67.4	66.8	69.7	67.0	67.8	**75.5**	**85.1**	**76.8**	**79.1**
•	•	◦	◦	71.1	11.9	1.2	28.1	83.9	56.7	29.0	56.5	85.5	52.6	25.3	54.5	84.3	55.3	24.2	54.6	**86.4**	**66.3**	**42.5**	**65.1**
◦	•	◦	•	47.3	17.2	0.6	21.7	80.8	53.4	28.3	54.2	80.8	48.6	25.2	51.5	82.2	57.2	30.7	56.7	**82.3**	**62.1**	**39.4**	**61.3**
•	◦	◦	•	74.8	17.7	0.8	31.1	86.0	58.7	28.0	57.6	86.3	50.7	25.2	54.1	**87.5**	59.7	34.6	60.6	85.9	**65.1**	**43.7**	**64.9**
•	◦	•	◦	68.4	41.4	53.8	54.5	83.3	67.6	68.0	73.0	**85.8**	72.5	70.4	76.2	**85.8**	72.9	70.3	76.2	**85.8**	**86.5**	**78.9**	**83.7**
•	•	•	◦	70.2	48.8	60.9	60.0	85.1	70.7	69.9	75.2	85.6	72.0	71.0	76.2	86.2	74.2	71.1	77.2	**86.6**	**87.1**	**78.8**	**84.1**
•	•	◦	•	75.2	18.7	1.0	31.6	87.0	61.0	33.4	60.5	86.1	52.5	25.8	54.8	**88.0**	61.5	34.1	61.2	86.5	**67.0**	**45.3**	**66.3**
•	◦	•	•	75.6	54.9	60.5	63.7	87.0	72.2	69.7	76.3	86.5	72.2	69.8	76.2	**88.6**	75.6	71.2	78.5	86.4	**86.7**	**78.8**	**83.9**
◦	•	•	•	44.2	46.6	55.1	48.6	82.1	70.7	69.7	74.2	81.1	69.5	68.5	73.0	**83.3**	75.3	71.1	76.6	82.6	**86.8**	**78.1**	**82.5**
•	•	•	•	73.8	55.3	61.1	63.4	87.6	73.4	70.8	77.3	86.3	71.8	69.9	76.0	**88.8**	76.4	71.7	79.0	86.5	**87.0**	**78.6**	**84.1**
Average				50.7	28.7	27.4	28.1	78.6	59.7	48.1	62.1	79.3	58.9	46.9	61.7	80.1	64.0	50.0	64.7	**82.4**	**75.0**	**60.0**	**72.5**

performance of the proposed method using three evaluation metrics: MSE, PSNR and SSIM. The comparison results are presented in Table 7. Overall, it can be observed the proposed network can obtain a stable generation performance no matter if any number of modalities are missing. For the Flair modality, the average generation accuracy is 0.033, 0.85 and 28.34 in terms of MSE, SSIM and PSNR, respectively, across all the missing situations; For the T1 modality, the average generation accuracy is 0.027, 0.86 and 25.62 in terms of MSE, SSIM and PSNR, respectively; For T1c modality, the average generation accuracy is 0.017, 0.86 and 32.15 in terms of MSE, SSIM and PSNR, respectively; For T2 modality, the average generation accuracy is 0.11, 0.80 and 22.98 in terms of MSE, SSIM and PSNR, respectively. The experimental results indicate that the proposed method can achieve good generation performance to help to recover the missing modalities. In addition, comparing the generation accuracy among the four modalities, it can be observed that the generation accuracy of Flair, T1 and T1c modalities is better than T2 modality. I explain that the timing of radiofrequency pulse sequences makes T2 modality different from other three modalities, for example, as is shown in Fig. 1, Cerebrospinal Fluid (CSF) is bright on T2 modality and dark on Flair, T1 and T1c modalities. Therefore, the generation accuracy of T2 modality is not better than the others.

5.5. Visualization of the segmentation and generation results

I also visualize the segmentation and generation results of the proposed method in Fig. 8. On the one hand, from the segmentation result, firstly, it can be observed that when only the FLAIR modality is available, it can detect most parts of brain tumor regions, indicating the important role of the FLAIR modality among these four MRI modalities. Secondly, when the T1c modality is integrated, the network can achieve good results. Significant improvements in DSC can be seen on both tumor core (+28.9%) and enhancing tumor (+99.5%) regions. Lastly, the T2 modality can further refine the results. On the other hand, from the generation results, it can be observed when only the FLAIR modality is available, it can obtain good generation results for both T1 and T2 modalities. When the T1c modality is included as the input (3rd and 4th row), the generated T1 and T2 modality can further highlight the tumor core region, resulting from the shared encoders with the segmentation network. In turn, the highlighted tumor regions can also boost the segmentation network to achieve better results.

From Fig. 9, in each row, it can be observed that with the help of the proposed strategies, the segmentation results can be gradually refined. In each column, with the increasing number

Table 7
Generation results of the proposed method evaluated by MSE, SSIM and PSNR on BraTS 2018 dataset. • denotes the included modality and ○ denotes the missing one. The colours distinguish the generation results of different modalities.

Modality				Generation Evaluation Metrics								
F	T1	T1c	T2	MSE ↓			SSIM ↑			PSNR ↑		
○	○	○	•	F: 0.048	T1: 0.03	T1c: 0.02	F: 0.83	T1: 0.85	T1c: 0.85	F: 26.62	T1: 25.65	T1c: 30.02
○	○	•	○	F: 0.028	T1: 0.025	T2: 0.097	F: 0.84	T1: 0.86	T2: 0.80	F: 28.37	T1: 27.33	T2: 22.78
○	•	○	○	F: 0.029	T1c: 0.019	T2: 0.12	F: 0.86	T1c: 0.87	T2: 0.80	F: 29.11	T1c: 31.16	T2: 22.96
•	○	○	○	T1: 0.034	T1c: 0.027	T2: 0.12	T1: 0.84	T1c: 0.84	T2: 0.80	T1: 24.99	T1c: 29.89	T2: 22.32
○	○	•	•	F: 0.038	T1: 0.025		F: 0.84	T1: 0.87		F: 27.75	T1: 21.55	
○	•	•	○	F: 0.025	T2: 0.097		F: 0.89	T2: 0.81		F: 29.46	T2: 23.25	
•	•	○	○	T1c: 0.018	T2: 0.11		T1c: 0.86	T2: 0.80		T1c: 39.0	T2: 23.1	
○	•	○	•	F: 0.034	T1c: 0.014		F: 0.84	T1c: 0.88		F: 28.38	T1c: 31.84	
•	○	○	•	T1: 0.026	T1c: 0.016		T1: 0.86	T1c: 0.86		T1: 26.16	T1c: 31.15	
•	○	•	○	T1: 0.026	T2: 0.11		T1: 0.85	T2: 0.80		T1: 26.47	T2: 23.13	
•	•	•	○	T2: 0.10			T2: 0.80			T2: 23.32		
•	•	○	•	T1c: 0.01			T1c: 0.88			T1c: 32.0		
•	○	•	•	T1: 0.024			T1: 0.86			T1: 27.21		
○	•	•	•	F: 0.032			F: 0.84			F: 28.69		

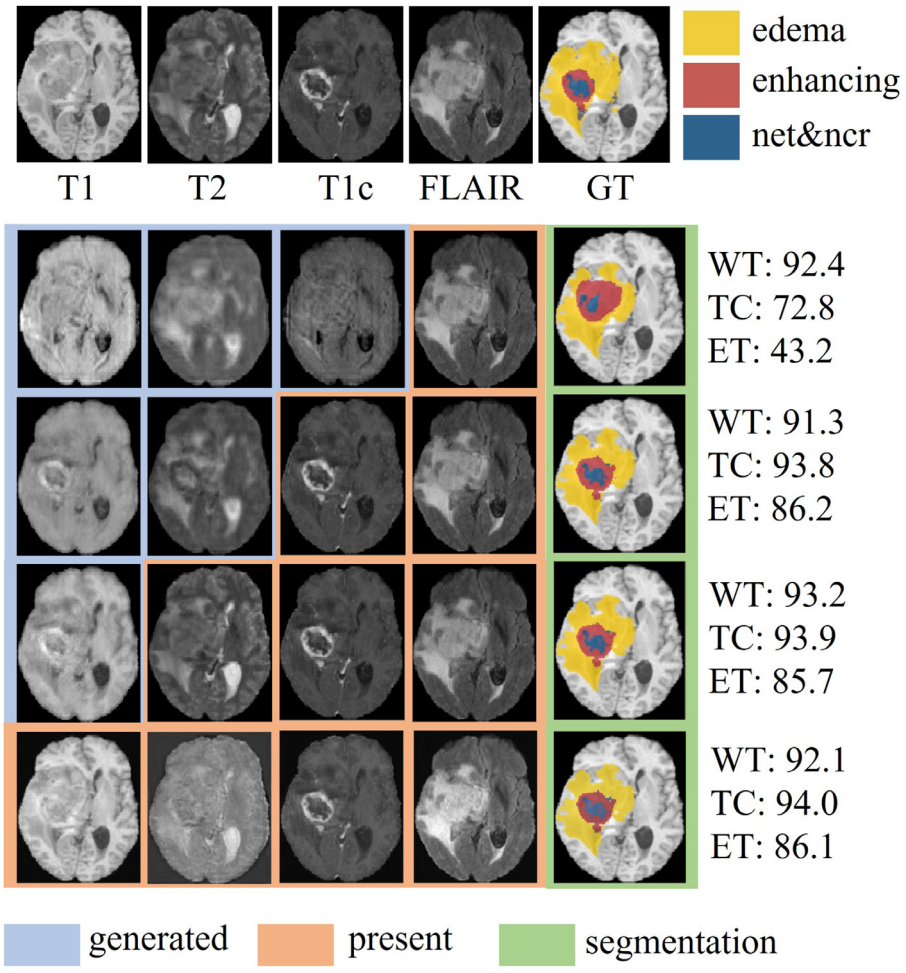

Fig. 8. The visualization of the segmentation and generation results of the proposed method. The first row presents the input modalities and the segmentation ground-truth. The next three rows present the generation and segmentation results in different missing modality situations. The last row presents the full modalities situation. DSC is denoted for each example.

of modalities, the segmentation results are improved progressively. Especially, from the second and third rows, it can be seen that integrating the T1c modality can significantly improve the segmentation results on both tumor core and enhancing tumor regions. This is consistent with the observation in Fig. 8, where the T1c modality is more sensitive to tumor core and enhancing tumor regions. To summarize, the visualization results show that the proposed method is able to achieve competitive segmentation when modalities are missing and can also generate the missing modalities at the same time.

5.6. Visualization of the feature maps

To further demonstrate the effectiveness of the proposed strategies, I visualize for various subsets of the strategies the feature maps extracted from the second-to-last layer in the segmenta-

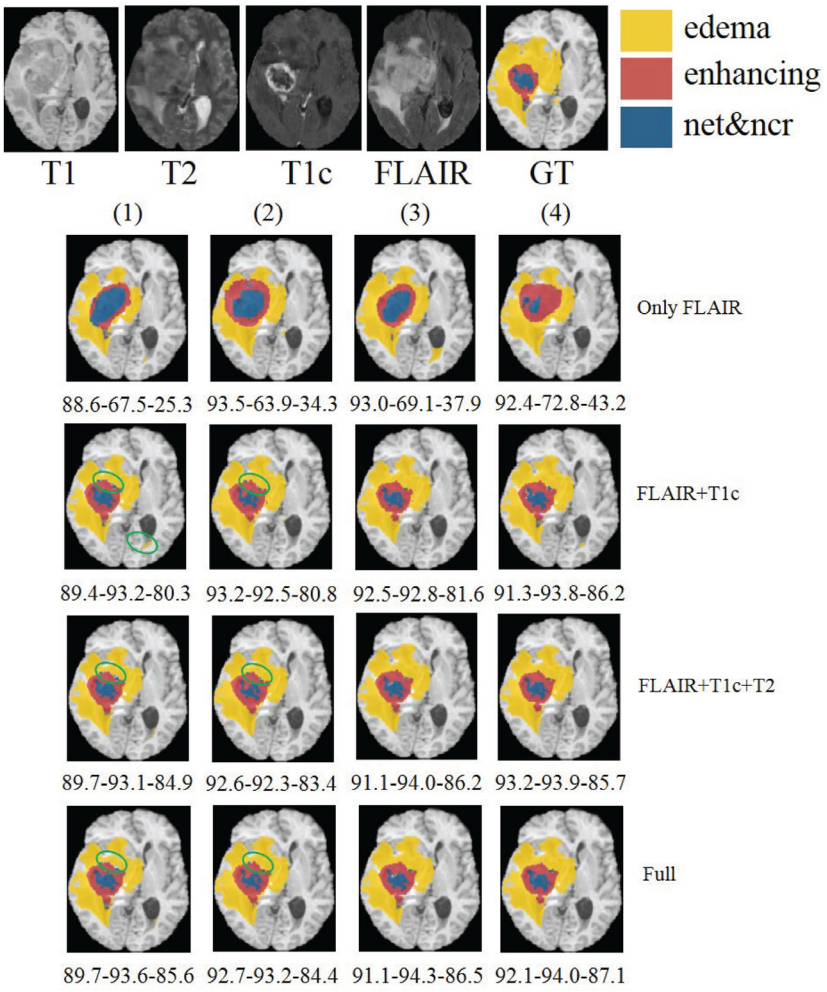

Fig. 9. The visualization of the segmentation results between different methods. The first row presents the input modalities and the segmentation ground-truth. The last four rows present the segmentation results in different missing modality situations between different methods. (1) Baseline, (2) + CMFM, (3) + CMFM + MSFM, (4) + CMFM + MSFM + SC-LFLM, DSC on the whole tumor, tumor core and enhancing tumor regions are denoted under each example. The green circle highlights the segmentation differences along columns. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

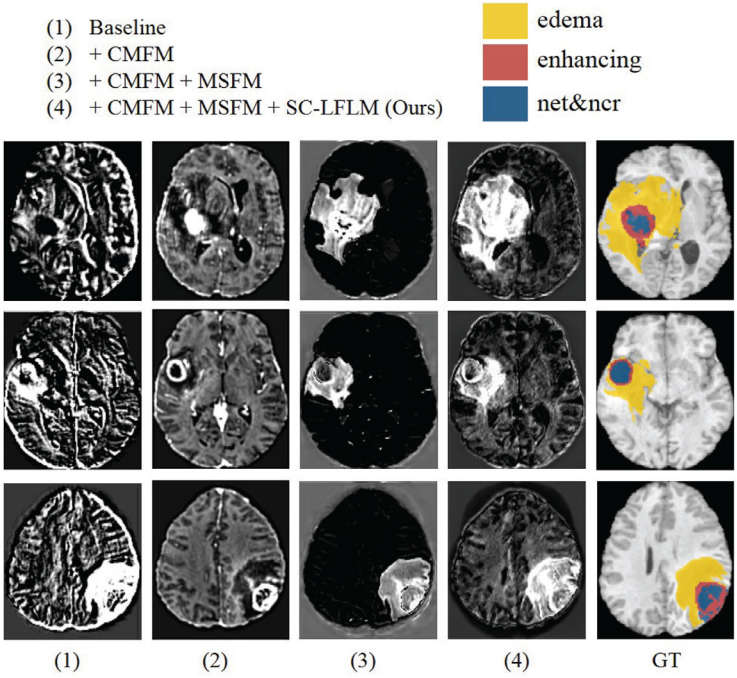

Fig. 10. The visualization of the feature maps between different methods.

tion decoder. From Fig. 10, it can be observed that with the help of the proposed strategies, the interested tumor regions become more obvious in the feature maps. Specifically, from the second and third columns, it can be observed employing the multi-scale feature information can significantly highlight the target tumor regions, especially for the edema region, aiding the network to learn more discriminative features for segmentation. Furthermore, based on the spatial consistency of multi-modalities, considering the latent relationships between different modalities can further obtain better feature information. Therefore, the comparison results in Fig. 10 demonstrate the effectiveness of the proposed method.

6. Discussion and conclusion

In this work, I proposed a novel multimodal feature fusion and latent feature learning guided deep neural network for brain tumor segmentation and missing modality recovery. Considering that the multi-modalities can provide complementary information for brain tumor segmentation, I propose to enhance the feature learning ability by introducing a multimodal feature fusion model, consisting of a cross-modality fusion module and a multi-scale fusion module. The cross-modality fusion module adopts the self-attention mechanism to learn non-local structures in the image. The multi-scale fusion module is proposed to capture multimodal spatial contextual feature information. Thanks to the two modules, the network can learn more rich features across multi-modalities. In addition, since the same tumor regions can be observed by different MR modalities, there is a spatial consistency between multi-modalities for the same patient. To this end, I proposed a spatial consistency-based latent feature learning module to learn the latent multimodal correlation, and also extract the relevant features to help segmentation. Furthermore, to compensate for the incomplete set of modalities, I propose to use multi-task learning to retrieve the missing modalities. Three generation evaluation metrics including MSE, PSNR and SSIM proved that the proposed network can achieve a stable generation performance with any number of missing modalities. Comprehensive experiments evaluated on BraTS 2018 dataset demonstrate that the proposed method can achieve superior segmentation results than the state-of-the-art methods. The proposed method is evaluated on BraTS 2018 dataset, while it can be generalized to other multimodal datasets. Besides, the proposed components such as the multimodal feature fusion model can be easily adapted to other neural network architectures and research fields.

However, there are some limitations in the work that inspire future directions. First, an encoder-decoder-based network is used for image generation, the potential future direction is to improve the generator architecture. For example, a Generative Adversarial Network (GAN) can be considered to generate high-quality images, as well as to further enhance segmentation accuracy. Second, the multi-task learning architecture can be improved to reduce training parameters and computational costs. Finally, other approaches to explore latent feature representations can be further investigated to improve the results.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62206084).

References

[1] Q.T. Ostrom, H. Gittleman, P. Liao, T. Vecchione-Koval, Y. Wolinsky, C. Kruchko, J.S. Barnholtz-Sloan, Cbtrus statistical report: primary brain and other central nervous system tumors diagnosed in the united states in 2010–2014, Neuro-Oncology 19 (suppl_5) (2017) v1–v88.
[2] E.B. Claus, K.M. Walsh, J.K. Wiencke, A.M. Molinaro, J.L. Wiemels, J.M. Schildkraut, M.L. Bondy, M. Berger, R. Jenkins, M. Wrensch, Survival and low-grade glioma: the emergence of genetic information, Neurosurg. Focus 38 (1) (2015) E6.
[3] S. Chen, C. Ding, M. Liu, Dual-force convolutional neural networks for accurate brain tumor segmentation, Pattern Recognit. 88 (2019) 90–100.
[4] T. Zhou, S. Ruan, S. Canu, A review: deep learning for medical image segmentation using multi-modality fusion, Array (2019) 100004.
[5] Y. Ding, W. Zheng, J. Geng, Z. Qin, K.-K.R. Choo, Z. Qin, X. Hou, Mvfusfra: a multi-view dynamic fusion framework for multimodal brain tumor segmentation, IEEE J. Biomed. Health Inform. (2021).
[6] M. Goetz, C. Weber, F. Binczyk, J. Polanska, R. Tarnawski, B. Bobek-Billewicz, U. Koethe, J. Kleesiek, B. Stieltjes, K.H. Maier-Hein, Dalsa: domain adaptation for supervised learning from sparsely annotated mr images, IEEE Trans. Med. Imag. 35 (1) (2015) 184–196.
[7] P. Singh, A type-2 neutrosophic-entropy-fusion based multiple thresholding method for the brain tumor tissue structures segmentation, Appl. Soft Comput. 103 (2021) 107119.
[8] Z.-Q. Zhao, P. Zheng, S.-t. Xu, X. Wu, Object detection with deep learning: a review, IEEE Trans. Neural Netw. Learn. Syst. 30 (11) (2019) 3212–3232.
[9] S.M. Marvasti-Zadeh, L. Cheng, H. Ghanei-Yakhdan, S. Kasaei, Deep learning for visual tracking: a comprehensive survey, IEEE Trans. Intell. Transp. Syst. (2021).
[10] Y.-J. Zheng, S.-L. Yu, Q. Song, Y.-J. Huang, W.-G. Sheng, S.-Y. Chen, Co-evolutionary fuzzy deep transfer learning for disaster relief demand forecasting, IEEE Trans. Emerg. Top. Comput. 10 (3) (2021) 1361–1373.
[11] R.J.S. Raj, S.J. Shobana, I.V. Pustokhina, D.A. Pustokhin, D. Gupta, K. Shankar, Optimal feature selection-based medical image classification using deep learning model in internet of medical things, IEEE Access 8 (2020) 58006–58017.
[12] K. Zhou, Y. Yang, T. Hospedales, T. Xiang, Deep domain-adversarial image generation for domain generalisation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 13025–13032.
[13] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, Y. Yu, Cross-modality deep feature learning for brain tumor segmentation, Pattern Recognit. 110 (2021) 107562.
[14] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, Y. Wang, Y. Yu, Exploring task structure for brain tumor segmentation from multi-modality mr images, IEEE Trans. Image Process. 29 (2020) 9032–9043.
[15] 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.
[16] A. Myronenko, 3d mri brain tumor segmentation using autoencoder regularization, in: International MICCAI Brainlesion Workshop, Springer, 2018, pp. 311–320.
[17] Z. Jiang, C. Ding, M. Liu, et al., Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. brainlesion: glioma, Multip. Scleros. Stroke Traumatic Brain Injur. (2020).
[18] F. Isensee, P.F. Jäger, P.M. Full, P. Vollmuth, K.H. Maier-Hein, nnu-net for brain tumor segmentation, in: International MICCAI Brainlesion Workshop, Springer, 2020, pp. 118–132.
[19] M. Islam, N. Wijethilake, H. Ren, Glioblastoma multiforme prognosis: mri missing modality generation, segmentation and radiogenomic survival prediction, Comput. Med. Imag. Graph. (2021) 101906.
[20] M. Havaei, N. Guizard, N. Chapados, Y. Bengio, Hemis: hetero-modal image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 469–477.
[21] K. Lau, J. Adler, J. Sjölund, A unified representation network for segmentation with missing modalities, arXiv preprint arXiv:1908.06683 (2019).
[22] A. Chartsias, T. Joyce, M.V. Giuffrida, S.A. Tsaftaris, Multimodal mr synthesis via modality-invariant latent representation, IEEE Trans. Med. Imag. 37 (3) (2017) 803–814.
[23] R. Dorent, S. Joutard, M. Modat, S. Ourselin, T. Vercauteren, Hetero-modal variational encoder-decoder for joint modality completion and segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 74–82.
[24] C. Chen, Q. Dou, Y. Jin, H. Chen, J. Qin, P.-A. Heng, Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 447–456.
[25] Y. Shen, M. Gao, Brain tumor segmentation on mri with missing modalities, in: International Conference on Information Processing in Medical Imaging, Springer, 2019, pp. 417–428.
[26] Y. Zhu, S. Wang, R. Lin, Y. Hu, Q. Chen, Brain tumor segmentation for missing modalities by supplementing missing features, in: 2021 IEEE 6th International

Conference on Cloud Computing and Big Data Analytics (ICCCBDA), IEEE, 2021, pp. 652–656.
[27] Y. Xia, L. Zhang, N. Ravikumar, R. Attar, S.K. Piechnik, S. Neubauer, S.E. Petersen, A.F. Frangi, Recovering from missing data in population imaging–cardiac mr image imputation via conditional generative adversarial nets, Med. Image Anal. 67 (2020) 101812.
[28] Y. Zhang, Q. Yang, A survey on multi-task learning, IEEE Trans. Knowl. Data Eng. (2021).
[29] T. He, J. Hu, Y. Song, J. Guo, Z. Yi, Multi-task learning for the segmentation of organs at risk with label dependence, Med. Image Anal. 61 (2020) 101666.
[30] H. Lin, J.D. Deng, D. Albers, F.W. Siebert, Helmet use detection of tracked motorcycles using cnn-based multi-task learning, IEEE Access 8 (2020) 162073–162084.
[31] Y. Zhang, Y. Bai, M. Ding, B. Ghanem, Multi-task generative adversarial network for detecting small objects in the wild, Int. J. Comput. Vis. (2020) 1–19.
[32] I.J. Jacob, Performance evaluation of caps-net based multitask learning architecture for text classification, J. Artif. Intell. 2 (01) (2020) 1–10.
[33] B. Bischke, P. Helber, J. Folz, D. Borth, A. Dengel, Multi-task learning for segmentation of building footprints with deep neural networks, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 1480–1484.
[34] H. Huang, G. Yang, W. Zhang, X. Xu, W. Yang, W. Jiang, X. Lai, A deep multi-task learning framework for brain tumor segmentation, Front. Oncol. 11 (2021).
[35] A. Foo, W. Hsu, M.L. Lee, G. Lim, T.Y. Wong, Multi-task learning for diabetic retinopathy grading and lesion segmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 13267–13272.
[36] A. Amyar, R. Modzelewski, H. Li, S. Ruan, Multi-task deep learning based ct imaging analysis for covid-19 pneumonia: classification and segmentation, Comput. Biol. Med. 126 (2020) 104037.
[37] P. Singh, S.S. Bose, Ambiguous d-means fusion clustering algorithm based on ambiguous set theory: special application in clustering of ct scan images of covid-19, Knowl. Based Syst. 231 (2021) 107432.
[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.
[39] T. Zhou, S. Ruan, P. Vera, S. Canu, A tri-attention fusion guided multi-modal segmentation network, Pattern Recognit. (2021) 108417.
[40] T. Zhou, S. Canu, S. Ruan, Fusion based on attention mechanism and context constraint for multi-modal brain tumor segmentation, Comput. Med. Imag. Graph. 86 (2020) 101811.
[41] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R.T. Shinohara, C. Berger, S.M. Ha, M. Rozycki, et al., Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge, arXiv preprint arXiv:1811.02629 (2018).

Tongxue Zhou received the Ph.D. degree in Computer Science from INSA Rouen, France, in 2022. She is currently a lecturer in the School of Information Science and Technology at Hangzhou Normal University, China. Her current research interests include medical image analysis, data fusion and deep learning.