frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 14 September 2022
DOI 10.3389/fnins.2022.1000587

OPEN ACCESS

EDITED BY
Xiangzhi Bai,
Beihang University, China

REVIEWED BY
Zhaoying Liu,
Beijing University of Technology, China
Junzhang Chen,
Beihang University, China

*CORRESPONDENCE
Xun Chen
xunchen@ustc.edu.cn

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 22 July 2022
ACCEPTED 18 August 2022
PUBLISHED 14 September 2022

CITATION
Liu Y, Mu F, Shi Y, Cheng J, Li C and Chen X (2022) Brain tumor segmentation in multimodal MRI via pixel-level and feature-level image fusion. *Front. Neurosci.* 16:1000587. doi: 10.3389/fnins.2022.1000587

Brain tumor segmentation in multimodal MRI *via* pixel-level and feature-level image fusion

Yu Liu[1,2], Fuhao Mu[1], Yu Shi[1], Juan Cheng[1,2], Chang Li[1,2] and Xun Chen[3]*

[1]Department of Biomedical Engineering, Hefei University of Technology, Hefei, China, [2]Anhui Province Key Laboratory of Measuring Theory and Precision Instrument, Hefei University of Technology, Hefei, China, [3]Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China

Brain tumor segmentation in multimodal MRI volumes is of great significance to disease diagnosis, treatment planning, survival prediction and other relevant tasks. However, most existing brain tumor segmentation methods fail to make sufficient use of multimodal information. The most common way is to simply stack the original multimodal images or their low-level features as the model input, and many methods treat each modality data with equal importance to a given segmentation target. In this paper, we introduce multimodal image fusion technique including both pixel-level fusion and feature-level fusion for brain tumor segmentation, aiming to achieve more sufficient and finer utilization of multimodal information. At the pixel level, we present a convolutional network named PIF-Net for 3D MR image fusion to enrich the input modalities of the segmentation model. The fused modalities can strengthen the association among different types of pathological information captured by multiple source modalities, leading to a modality enhancement effect. At the feature level, we design an attention-based modality selection feature fusion (MSFF) module for multimodal feature refinement to address the difference among multiple modalities for a given segmentation target. A two-stage brain tumor segmentation framework is accordingly proposed based on the above components and the popular V-Net model. Experiments are conducted on the BraTS 2019 and BraTS 2020 benchmarks. The results demonstrate that the proposed components on both pixel-level and feature-level fusion can effectively improve the segmentation accuracy of brain tumors.

KEYWORDS

brain tumor segmentation, medical image fusion, pixel-level fusion, feature-level fusion, convolutional neural networks

1. Introduction

Automatically and accurately segmenting brain tumor areas from multimodal magnetic resonance imaging (MRI) scans can provide crucial information about tumors including shape, volume, and localization. Based on these information, quantitative assessment of lesions can be carried out, which is of great significance to disease

diagnosis, treatment planning, survival prediction, and other relevant tasks. Most existing brain tumor segmentation studies are concentrating on gliomas since they are the most common brain tumors in adults. However, due to the factors like the variety of tumor size, shape and position, the fuzzy boundaries, and the difference in intensity distribution of MRI data obtained by different devices, the accurate segmentation of brain tumors is always a very challenging task (Zhao et al., 2018).

Owing to the good ability in capturing high-resolution anatomic structure of tissues, MRI is mostly used in brain tumor segmentation. Commonly-used MRI modalities for brain tumor segmentation include T1-weighted (T1), contrast-enhanced T1-weighted (T1c), T2-weighted (T2), and fluid attenuated inversion recovery (Flair). Figure 1 gives an example of multimodal MRI volumes for brain tumor segmentation, which comes from the dataset released by the Brain Tumor Segmentation (BraTS) challenge (Menze et al., 2015), an annual event held by the Medical Image Computing and Computer Assisted Intervention (MICCAI). The segmentation label (i.e., ground truth) provided by physicians is also shown in Figure 1. The green, red, and yellow regions indicate edema (ED), necrosis and non-enhancing tumor (NCR/NET), and enhancing tumor (ET), respectively. In the BraTS challenge, the segmentation performance is evaluated on three partially overlapping sub-regions of tumors, namely, whole tumor (WT), tumor core (TC), and enhancing tumor (ET). The WT is the union of ED, NCR/NET, and ET, while the TC includes NCR/NET and ET. We can see from Figure 1 that different pathological features of tumors are captured by MRI data of different modalities.

In recent years, various brain tumor segmentation methods have been proposed. Traditional image segmentation methods based on threshold, region, and pixel clustering are difficult to achieve good results in this task due to its high complexity as mentioned above (Liu et al., 2014). The performance of machine learning approaches based on hand-crafted features and classifiers like support vector machines and random forests is still limited in most cases. In the last few years, deep learning-based methods have emerged as the trend in this field due to their obvious advantages on segmentation accuracy (Bakas et al., 2018). Some methods adopt a 2D or 3D patch-based manner, in which convolutional networks are applied to predict the class of the center voxel (Havaei et al., 2017; Kamnitsas et al., 2017; Zhao et al., 2018). However, these methods tend to ignore the correlation among different patches within a large receptive field. To better address the global contextual information, the encoder-decoder architectures represented by U-Net (Ronneberger et al., 2015) and V-Net (Milletari et al., 2016) have become more and more popular in brain tumor segmentation (Wang et al., 2017; Li et al., 2019a; Zhang et al., 2020a; Zhou et al., 2020).

As brain tumor segmentation in MRI is essentially a multimodal image segmentation problem, the joint utilization of multimodal information plays a critical role in this task (Zhang et al., 2022). However, we argue that most existing methods do not pay enough attention to this issue and the utilization of multimodal information is not sufficient. In existing brain tumor segmentation methods, the most common way of using multimodal MR images is to simply stack them or their low-level features as the model input (Cao et al., 2021; Chen et al., 2021; Valanarasu et al., 2021; Wang et al., 2021; Zhang et al., 2021b). In addition, as mentioned above, MR images with different modalities reflect different pathological features (Chen et al., 2021; Wang et al., 2021), so their importance to a given segmentation target should be different. However, many methods fail to take this difference into consideration in their segmentation models and there is a lack of refinement for multimodal features, which will have an adverse effect on the segmentation performance.

In this paper, we address the above problems *via* the multimodal image fusion technique at both the pixel level and the feature level. For one thing, we adopt pixel-level image fusion to enrich the input modalities of the segmentation model and the fused modalities can strengthen the association among different types of pathological information captured by multiple source modalities. For another, we embed an attention-based feature fusion module into the segmentation network to refine multimodal features for better segmentation performance.

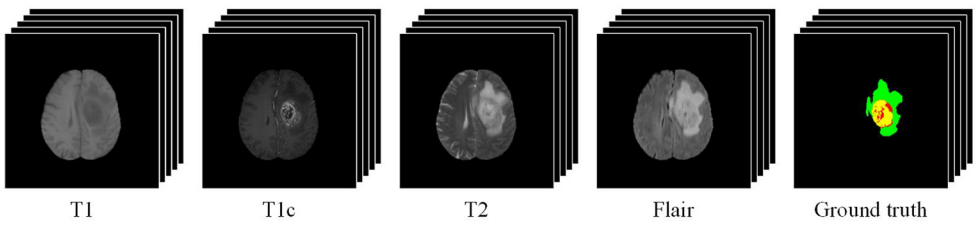

FIGURE 1
An example of multimodal MRI volumes for brain tumor segmentation. The green, red, and yellow regions in the ground truth indicate edema (ED), non-enhancing tumor and necrosis (NCR/NET), and enhancing tumor (ET), respectively.

Specifically, the main contributions of this work are summarized into four points:

1. To make use of multimodal information more sufficiently for brain tumor segmentation, we introduce the multimodal image fusion technique including both pixel-level fusion and feature-level fusion into the segmentation task.
2. We present a pixel-level image fusion network (PIF-Net) to fuse 3D multimodal MR images, aiming to enrich the input modalities of the segmentation model. This is actually a modality enhancement approach since the fused modalities obtained by the PIF-Net can effectively combine the pathological information from multiple source modalities.
3. To address the difference among multiple modalities for a given segmentation target, we design an attention-based modality selection feature fusion (MSFF) module for multimodal feature refinement and it is embedded into the segmentation network for performance improvement.
4. We propose a two-stage brain tumor segmentation framework based on the PIF-Net, the MSFF module and the V-Net. Experimental results on the BraTS 2019 and BraTS 2020 benchmarks demonstrate the effectiveness of the proposed pixel-level and feature-level fusion approaches for brain tumor segmentation.

The rest of this paper is organized as follows. Section 2 introduces the related works. In Section 3, the proposed method is presented in detail. The experimental results and discussion are given in Section 4. Finally, we conclude the paper in Section 5.

2. Related work

2.1. Brain tumor segmentation

Many automatic brain tumor segmentation methods have been proposed in recent years. They can be roughly divided into two categories (Havaei et al., 2017): the generative model-based methods and the discriminative model-based methods. The generative model-based methods require domain-specific prior knowledge about the appearance characteristics of tumorous and healthy tissues, but they are challenging to characterize due to the complexity of brain tissues. The discriminative model-based methods treat brain tumor segmentation as a pattern classification problem for the voxels in MRI volumes and they have become the mainstream in this field owing to the rapid development of machine learning techniques. Popular hand-crafted features used in brain tumor segmentation include local histograms (Goetz et al., 2014), structure tensor eigenvalues (Kleesiek et al., 2014), texture features (Subbanna et al., 2013), and so on, while typical shallow learning models such as support vector machines and random forests are frequently adopted in brain tumor segmentation (Bauer et al., 2011; Meier et al., 2014; Pinto et al., 2015).

In the last few years, deep learning has rapidly achieved the dominance in brain tumor segmentation owing to the significantly improved performance. Some early methods adopt a patch-based classification manner by utilizing convolutional networks to predict the class of the center voxel of a 2D or 3D image patch. Havaei et al. (2017) proposed a two-pathway architecture to extract features with 2D convolutional kernels of different sizes. They also explored three cascade architectures in which the output of the first network with larger input size is supplemented as an additional source for the second network to extract information of multiple scales simultaneously. The DeepMedic (Kamnitsas et al., 2017), a well-known 3D brain tumor segmentation model proposed by Kamnitsas et al., also adopts a dual pathway architecture that uses patches of different sizes as the network input, aiming to incorporate both local and larger contextual information. In addition, the dense training scheme is employed in Kamnitsas et al. (2017) to address the relationship among neighboring patches. Zhao et al. (2018) integrated fully convolutional neural networks (FCNNs) and the conditional random field (CRF) into a unified framework for brain tumor segmentation. In their method, features are also extracted from receptive fields of different sizes.

The above patch-based classification methods can't fully consider the correlation among neighboring patches and the range of the receptive field is always limited, although some improved strategies are adopted. To address this problem, the encoder-decoder semantic segmentation architectures such as U-Net (Ronneberger et al., 2015), 3D U-Net (Çiçek et al., 2016), and V-Net (Milletari et al., 2016) have become more and more popular in brain tumor segmentation. Myronenko (2018) proposed a segmentation method that won the first place in the BraTS 2018 challenge by adding an variational auto-encoder (VAE) branch into an encoder-decoder architecture to obtain an additional regularization to the encoder part. To alleviate the issue of class imbalance, some methods apply a cascaded architecture to decompose the original multi-label segmentation problem into multiple binary segmentation sub-problems. Wang et al. (2017) cascaded three CNNs to realize the segmentation of three tumor areas including WT, TC and ET. Zhang et al. (2020a) proposed a task-structured brain tumor segmentation network to address the task-modality and task-task relationship simultaneously. Zhou et al. (2020) proposed a one-pass multi-task network with cross-task guided attention for brain tumor segmentation, which integrates the multiple segmentation sub-tasks into one deep model. Li et al. (2019a) proposed a multi-step cascaded network that takes the hierarchical topology of the brain tumor sub-structures into account and segments the sub-structures from coarse to fine.

However, it is worth noting that current study on brain tumor segmentation does not pay enough attention to the joint utilization of multimodal MR images, which is in fact a key

issue in this multimodal image segmentation task (Zhang et al., 2022). The most common way of using multimodal MR images is to simply stack them or their low-level features as the model input (Cao et al., 2021; Chen et al., 2021; Valanarasu et al., 2021; Wang et al., 2021; Zhang et al., 2021b). In addition, many methods treat each modality data with equal importance to a given segmentation target (Chen et al., 2021; Wang et al., 2021). These factors motivate us to introduce image fusion technique including both pixel-level fusion and feature-level fusion into the brain tumor segmentation framework for better performance.

2.2. Pixel-level medical image fusion

The purpose of pixel-level medical image fusion is to integrate the complementary information contained in multimodal medical images by generating a composite fused image, which is expected to be more suitable for human or machine perception. A variety of medical image fusion methods have been proposed over the past few decades and most of them are developed under a "decomposition-fusion-reconstruction" three-phase framework (Li et al., 2017; Liu et al., 2020b). Specifically, the source images are first decomposed into a transform domain and the decomposed coefficients from different source images are then fused. The fused image is finally reconstructed based on the fused coefficients. Multi-scale transform (MST) and sparse representation (SR) are two main categories of image decomposition that are widely used in medical image fusion (Liu et al., 2015, 2016, 2019, 2021; Du et al., 2016; Yang et al., 2016; Li et al., 2017; Zhang et al., 2018; Zhu et al., 2018; Yin et al., 2019).

However, most previous works in medical image fusion focus on the 2D image fusion problem, while methods for 3D image fusion were rarely studied (Yin, 2018). Using 2D fusion methods to tackle 3D medical images slice by slice independently neglects the correlation among adjacent slices and thereby tends to lose spatial contextual information of volumetric data. Wang et al. (2014) proposed a 3D multimodal medical image fusion method based on the 3D discrete shearlet transform (3D-DST) and designed a global-to-local strategy to fuse the decomposed coefficients. Yin (2018) introduced the tensor sparse representation (TSR), which is a high-dimensional extension of 2D SR, for 3D medical image fusion. Nevertheless, in these methods, the source images are treated equally in the fusion framework with identical decomposition approach and isotropic fusion strategy. As a result, the characteristics of different source modalities are not fully considered, leading to the loss of important modality information.

Recently, deep learning has emerged as an active direction in the field of image fusion (Liu et al., 2018; Zhang et al., 2021a) and some medical image fusion methods based on deep learning models like CNNs and generative adversarial networks (GANs) have been proposed (Liu et al., 2017, 2022; Liang et al., 2019; Ma et al., 2020a, 2022; Zhang et al., 2020b; Tang et al., 2021; Xu and Ma, 2021; Xu et al., 2022). By optimizing the loss functions that are specially designed based on the characteristics of source modalities, the deep learning-based methods have advantages over conventional MST-based and SR-based fusion methods on preserving modality information. However, the above deep learning-based methods are generally developed for 2D image fusion. In this work, we present a CNN-based 3D medical image fusion approach and introduce it for brain tumor segmentation by enriching the input modalities. In fact, current study on pixel-level medical image fusion is mostly devoted to pursuing good visual quality for physician observation and high evaluation results on objective metrics of image fusion, while very few study focuses on the application of image fusion to some specific clinical machine vision problems such as classification, detection and segmentation. Therefore, this work is also of high significance from the viewpoint of medical image fusion.

3. The proposed method

3.1. Overview

Figure 2 shows the schematic diagram of the proposed brain tumor segmentation framework. It consists of two stages to achieve the segmentation result of WT, TC, and ET areas. The two stages share a similar architecture that is composed of a PIF-Net to enrich the input modalities of the segmentation model *via* pixel-level fusion, an MSFF module to refine the mutlimodal features *via* feature-level fusion, and a V-Net (Milletari et al., 2016) with the encoder-decoder structure to obtain the segmentation result. The target of the first stage is to segment the WT area, while the second stage aims to identify the TC and ET areas. Since the TC and ET areas are included in the WT area, the segmentation result of the first stage is used to locate the input region of the second stage, which is helpful to alleviate the class imbalance issue. The sliding window-based approach introduced in Lyu and Shu (2020) is adopted to determine the input region of the second stage, namely, the window that contains the maximum number of tumor voxels is selected. In addition, considering that the peritumoral edema are mainly highlighted in T2 and Flair modalities, we only use T2 and Flair as the input source modalities in the first stage. The PIF-Net is used to generate their fused modality, which is denoted as T2-Flair. These three modalities (i.e., T2, Flair and T2-Flair) are fed together to the subsequent MSFF module in the first stage. In the second stage, all the four source modalities (i.e., T1, T1c, T2, and Flair) are adopted as the original input. The PIF-Net is applied to obtain two additional fused modalities, which are the fusion of T1c and T2 (denoted as T1c-T2), and the fusion of T1c and Flair (denoted as T1c-Flair). We mainly choose the T1c modality for

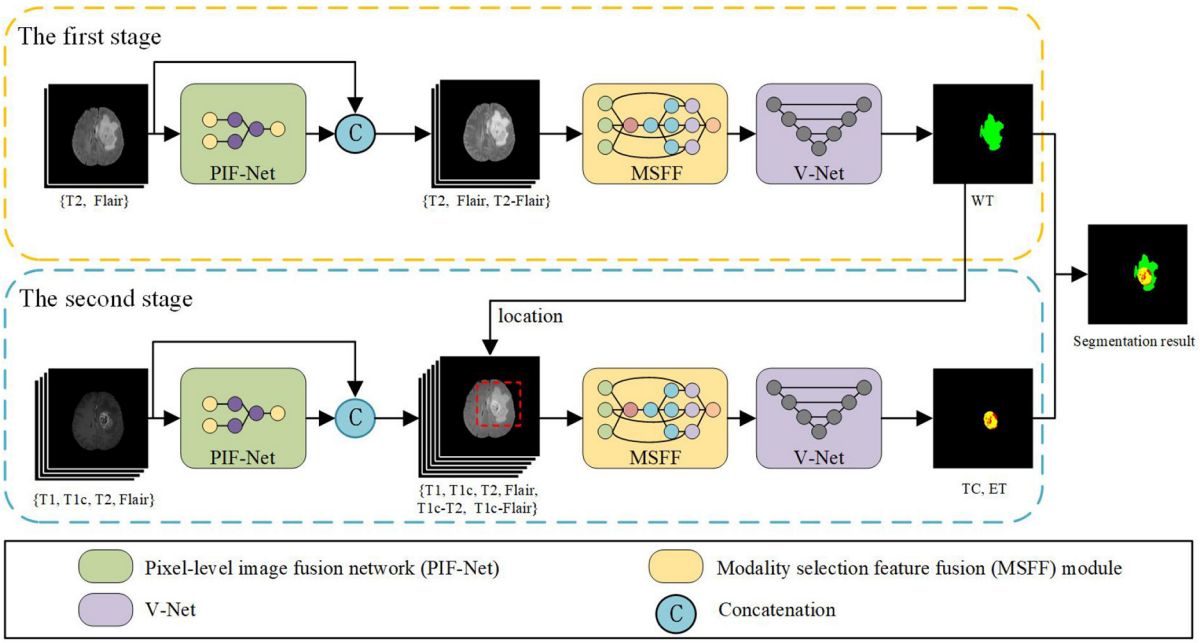

FIGURE 2
The schematic diagram of the proposed brain tumor segmentation framework.

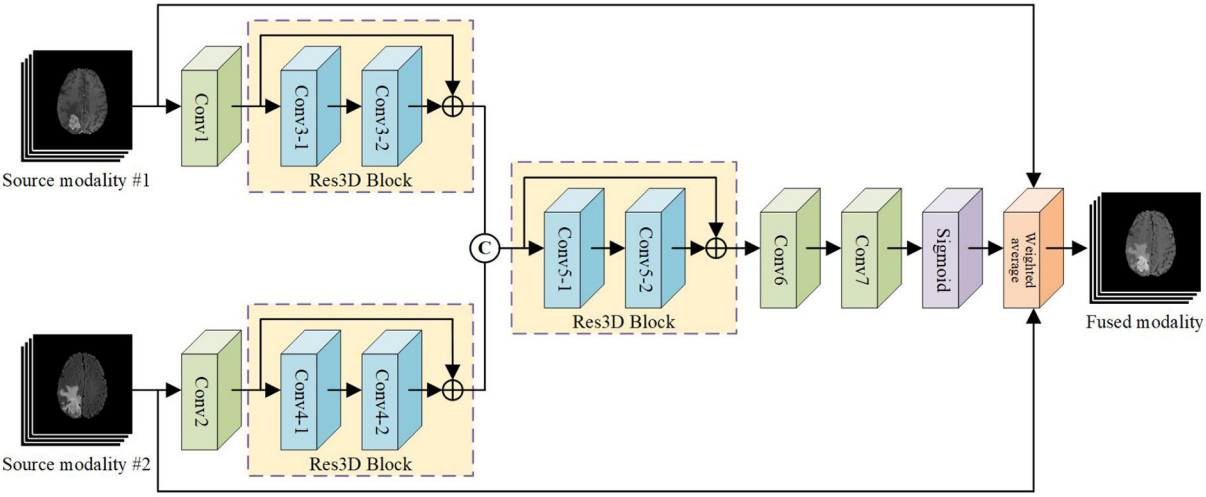

FIGURE 3
The architecture of our PIF-Net for 3D multimodal MR image fusion.

fusion because it is known to be very effective in detecting the TC and ET areas. By contrast, the T1 modality provides relatively less information for segmenting brain tumors and it generally plays an auxiliary role in this task (Bakas et al., 2018; Ma and Yang, 2018). Thus, the input of the MSFF module in the second stage contains six modalities in total. The final segmentation result is achieved by combining the results obtained at two stages together.

3.2. PIF-Net

Considering the high computational cost and memory usage of 3D convolutional networks, we design a relatively plain network architecture as shown in Figure 3 for 3D pixel-level image fusion. Note that this is likely to be the first work on CNN-based 3D medical image fusion to our knowledge, as mentioned in Section 2.2. The PIF-Net contains two branches for feature

TABLE 1 Detailed parameter configuration of the PIF-Net.

Layer	K_s	S_s	P_s	I_c	O_c	A
Conv1	$3 \times 3 \times 3$	1	1	1	32	ReLU
Conv2	$3 \times 3 \times 3$	1	1	1	32	ReLU
Conv3-1	$3 \times 3 \times 3$	1	1	32	32	ReLU
Conv3-2	$3 \times 3 \times 3$	1	1	32	32	/
Addition	/	/	/	32	32	ReLU
Conv4-1	$3 \times 3 \times 3$	1	1	32	32	ReLU
Conv4-2	$3 \times 3 \times 3$	1	1	32	32	/
Addition	/	/	/	32	32	ReLU
Conv5-1	$3 \times 3 \times 3$	1	1	64	64	ReLU
Conv5-2	$3 \times 3 \times 3$	1	1	64	64	/
Addition	/	/	/	64	64	ReLU
Conv6	$3 \times 3 \times 3$	1	1	64	32	/
Conv7	$3 \times 3 \times 3$	1	1	32	1	/
Sigmoid	/	/	/	1	1	/
Weighted average	/	/	/	1	1	/

K_s, S_s, P_s, I_c, O_c, and A denote the kernel size, stride, padding size, number of input channels, number of output channels, and activation operation, respectively.

extraction from two source modalities. Each branch is composed of a $3 \times 3 \times 3$ convolutional layer and a 3D residual (denoted as Res3D) block that contains two $3 \times 3 \times 3$ convolutional layers using the skip connection. The feature maps obtained from two branches are then concatenated and fed to another Res3D block. Two $3 \times 3 \times 3$ convolutional layers are further applied to reduce the number of channels to 1 and a sigmoid operation is conducted to reconstruct a weight mask. Finally, the fused modality is reconstructed by performing the weighted average calculation based on the mask and source images. It is worth noting that the fused image can also be reconstructed directly from the fused feature maps without using a weight mask. However, since the voxels in the meaningless background regions have zero-valued intensity in each source modality, a direct regression tends to cause inappropriate non-zero predictions in these regions, which will affect the fusion quality. The voxel-wise weighted average strategy adopted can effectively avoid this problem and we experimentally found that it can produce good fusion results. The detailed parameter configuration of the network architecture is given in Table 1.

The definition of loss function is a key issue in deep learning-based image fusion methods as it determines the preservation of modality information from source images. In this work, the loss function of our PIF-Net is formulated as

$$L_{pif} = L_{pixel} + \alpha L_{ssim}, \tag{1}$$

where L_{pixel} and L_{ssim} indicate the pixel loss and the structural similarity loss, respectively. α is the regularization parameter that balances these two terms, and it is experimentally set to 450 in our method. The pixel loss is designed to preserve the intensity information, which is often related to the lesions(e.g., edema) that have very high or low intensity in some MRI modalities. It is defined as

$$L_{pixel} = ||\mathbf{F} - \mathbf{S}_1||_F^2 + \beta||\mathbf{F} - \mathbf{S}_2||_F^2, \tag{2}$$

where $\mathbf{S}_1$ and $\mathbf{S}_2$ denote the source images, and $\mathbf{F}$ denotes the fused image. β is the trade-off parameter and $||\cdot||_F^2$ denotes the tensor Frobenius norm. The structural similarity loss is adopted to extract anatomic structure information from source images and it is defined as

$$L_{ssim} = \gamma(1 - \mathrm{SSIM}(\mathbf{F}, \mathbf{S}_1)) + (1 - \mathrm{SSIM}(\mathbf{F}, \mathbf{S}_2)), \tag{3}$$

where $\mathrm{SSIM}(\cdot, \cdot)$ represents the 3D structural similarity measure and γ is the trade-off parameter.

The parameters β and γ are set according to the specific characteristics of fusion problems. In the first stage, for the fusion of T2 and Flair images, β and γ are both set to 1 since these two modalities have relatively similar pathological and structural information. In the second stage, let $\mathbf{S}_1$ and $\mathbf{S}_2$ denote the T1c and T2/Flair images, respectively. Considering that the T2/Flair image contains more lesion information regarding the edema area, we increase the weight of T2/Flair images in L_{pixel}. Meanwhile, since the T1c image captures more tissue structures in the TC and ET areas, a larger weight is assigned to the T1c image in L_{ssim}. In our method, we set both β and γ to 2 for the fusion of T1c and T2/Flair images.

The PIF-Net is trained based on the training set released by the BraTS challenge 2019. The training set contains 335 cases of multimodal MRI volumes and four modalities (i.e., T1, T1c, T2, and Flair) are provided in each case. The original volumes of size $155 \times 240 \times 240$ are cropped into patches of size $80 \times 80 \times 80$ by the sliding window technique to enlarge the scale of the training set. The learning rate is fixed as 10^{-4} during the training process and the Adam optimizer is adopted to train the network. Figure 4 shows an example of fusion results obtained by the PIF-Net. The results of two representative 3D medical image fusion methods 3D-DST (Wang et al., 2014) and TSR (Yin, 2018) are also provided for comparison. The results of T2 and Flair fusion and T1c and Flair fusion are given at the first and second rows in Figure 4, respectively. It can be seen that the PIF-Net achieves higher fusion quality than the other two methods on the tumor areas, especially for the T1c and Flair fusion, in which the 3D-DST and TSR methods fail in preserving the edema information contained in the Flair images well, while the PIF-Net simultaneously preserve important modality information from both two source images.

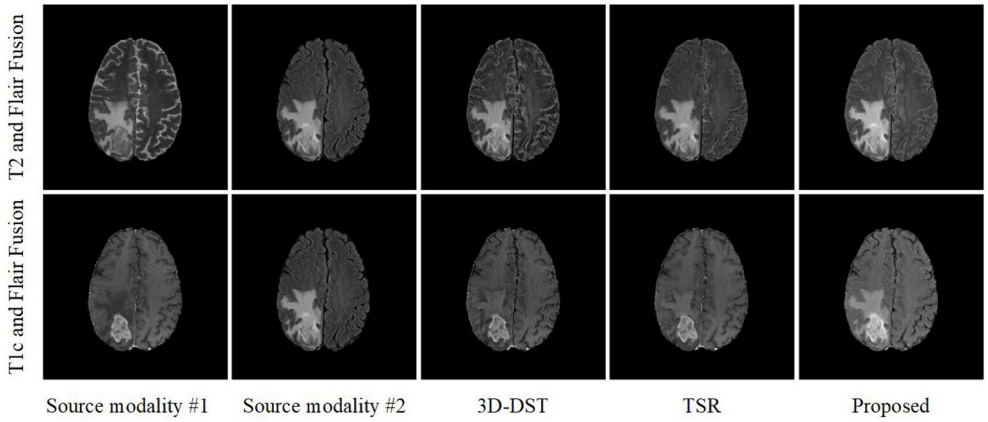

FIGURE 4
An example of fusion results obtained by different 3D medical image fusion methods.

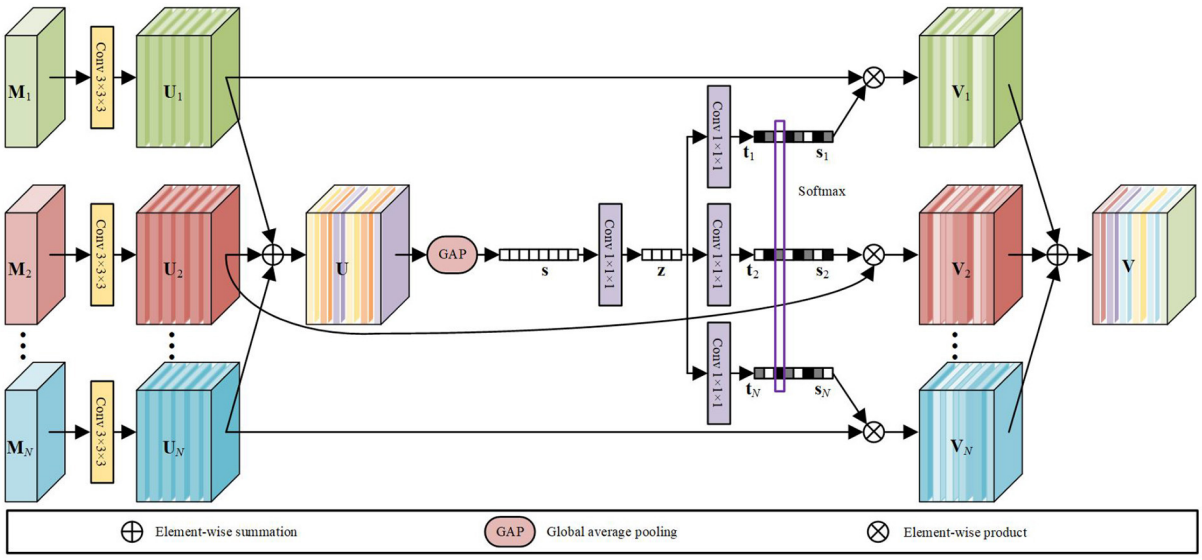

FIGURE 5
The architecture of our MSFF module for multimodal feature refinement.

3.3. MSFF module

The MSFF module is designed to refine the features extracted from multimodal MRI volumes for subsequent segmentation. Inspired the selective kernel network (SKNet) for multi-scale feature extraction (Li et al., 2019b), an attention-based feature fusion module is presented to adaptively adjust the weights of the features from different modalities. The architecture of our MSFF module is shown in Figure 5. Let $\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_N \in \mathbb{R}^{L\times H\times W\times 1}$ denote the input multimodal MRI volumes that involve both the original source modalities and the fused modalities obtained by the PIF-Net, where N is total number of input modalities. A $3 \times 3 \times 3$ convolutional layer is firstly performed on each input volume for feature extraction. The obtained features are denoted as $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_N \in \mathbb{R}^{L\times H\times W\times C}$, where $L \times H \times W$ denotes the size of the 3D feature map and C denotes the number of feature maps. In our method, C is set to 16. The features from different sources are firstly merged *via* an element-wise summation as

$$\mathbf{U} = \sum_{i=1}^{N} \mathbf{U}_i. \tag{4}$$

Then, we embed the global information by a channel-wise global average pooling (GAP) operation to get a feature vector

$\mathbf{s} \in \mathbb{R}^{1\times1\times1\times C}$. Specifically, the c-th element of $\mathbf{s}$ is calculated as

$$s_c = \Phi_{GAP}(\mathbf{U}_c) = \frac{1}{L \times H \times W} \sum_{i=1}^{L} \sum_{j=1}^{H} \sum_{k=1}^{W} \mathbf{U}_c(i, j, k). \quad (5)$$

Further, a compact feature $\mathbf{z} \in \mathbb{R}^{1\times1\times1\times C/r}$ is generated by a $1 \times 1 \times 1$ convolutional layer for channel reduction, which is actually equivalent to a fully connected layer. The ratio factor r is set to 4 in our model. Next, we adopt N parallel channel up-scaling convolutions with kernel size of $1 \times 1 \times 1$ to reconstruct N C-dimensional vectors $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_N \in \mathbb{R}^{1\times1\times1\times C}$. This is actually the excitation operation used in the SENet (Hu et al., 2018). Subsequently, a channel-wise softmax calculation is performed on each element across all the N vectors (indicated by the purple frame) to obtain the attention vectors $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N \in \mathbb{R}^{1\times1\times1\times C}$. Specifically, the c-th element of $\mathbf{s}_i$ is calculated as

$$s_{i,c} = \frac{e^{t_{i,c}}}{\sum_{j=1}^{N} e^{t_{j,c}}}, \quad (6)$$

where $t_{i,c}$ denotes the c-th element of $\mathbf{t}_i$, $i \in \{1, 2, \ldots, N\}$, $c \in \{1, 2, \ldots, C\}$.

Finally, the fused feature $\mathbf{V} \in \mathbb{R}^{L\times H\times W\times C}$ is calculated by a channel-wise weighted average over the source features using the attention weights as

$$\mathbf{V} = \sum_{i=1}^{N} \mathbf{s}_i \cdot \mathbf{U}_i. \quad (7)$$

According to a recent survey on attention mechanism (Guo et al., 2022), the attention mechanism used in our MSFF module belongs to the branch attention, which can be viewed as a dynamic branch selection mechanism and typically used in a multi-branch architecture. In the proposed method, to be more specific, the attention mechanism can be regarded as a kind of modality attention, aiming to extract features from multimodal MR images more effectively.

3.4. Segmentation loss

The loss function used for training the segmentation model is defined as

$$L_{seg} = L_{dice} + \lambda L_{bce}, \quad (8)$$

where L_{dice} and L_{bce} denote the dice loss and the binary cross entropy (BCE) loss, respectively, as

$$L_{dice} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2 + \varepsilon}, \quad (9)$$

$$L_{bce} = -\frac{1}{N} \sum_{i=1}^{N} [g_i \log p_i + (1 - g_i) \log(1 - p_i)], \quad (10)$$

where $g_i \in G$ is the ground truth binary volume, $p_i \in P$ is the network prediction, and N denotes the number of voxels. The parameter ε is a small constant to avoid dividing by 0. The Dice loss is known to be capable of alleviating the class imbalance issue (Milletari et al., 2016), while the BCE is the mostly used loss function for binary classification or segmentation. In brain tumor segmentation, the union of these two losses is a common way as it can combine their complementary advantages. The parameter λ controls the trade-off between these two losses and it is experimentally set to 0.5 in our method.

4. Experimental results and discussion

4.1. Data and implementation details

The BraTS 2019 and BraTS 2020 benchmarks (Menze et al., 2015) are adopted to demonstrate the effectiveness of the proposed method. The multimodal MRI data in a BraTS benchmark is divided into three parts: a training set, a validation set and a testing set. Only the training set releases the segmentation label (i.e., ground truth) annotated by experts to the public. The validation set is used to adjust model training and the MRI data is available, but the label is not provided. Users must upload their segmentation results to the organizer's sever at https://ipp.cbica.upenn.edu/ to obtain the evaluation results. Both data and label in the testing set are not available to users. In our experiments, just as most previous studies in this field, we adopt the training set for model training and validation, while use the validation set for performance evaluation. In particular, the BraTS training set is further divided into two parts: 80% samples are used for network training and the remaining 20% samples are used as a validation set to guide the training process. The BraTS 2019 training dataset includes 335 cases, while BraTS 2020 has a larger one comprising 369 cases. These multimodal MRI data have been skull-striped, re-sampled, and co-registered. Each case contains MRI data of four modalities (i.e., T1, T1c, T2, and Flair) and each volume is of size $155 \times 240 \times 240$.

For data pre-processing and augmentation, the popular z-score normalization approach is applied to each MRI volume, namely, the data is subtracted by the mean and divided by the standard deviation of the non-zero region. The training volume is randomly cropped into patches of size $128 \times 192 \times 160$ before fed to the network in the first stage. For each volume, the patch of size $128\times128\times128$ that contains maximum tumor voxels is used for training in the second stage. Moreover, in both two stages, the intensity of each volume is randomly shifted by a value in $[-0.1\sigma, 0.1\sigma]$ (σ denotes the standard deviation) and randomly

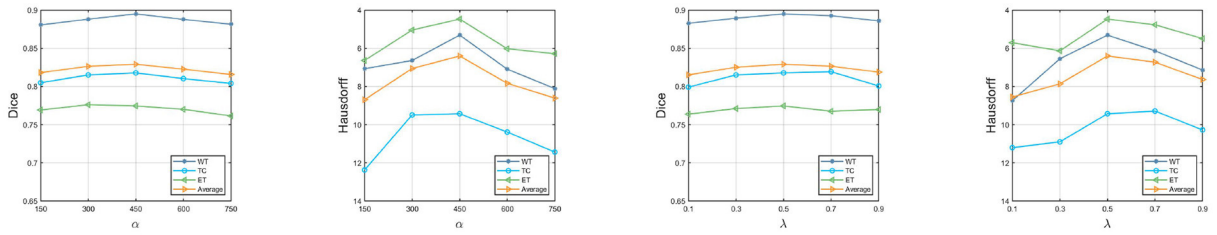

FIGURE 6
Impact of the parameters α and λ on the model performance.

scaled by a factor in $[0.9, 1.1]$. In addition, a random flipping along each axis is applied with a probability of 50%.

Our network is implemented in PyTorch and trained on two NVIDIA TITAN RTX GPUs. The Adam optimizer is used for updating weights. The learning rate is progressively decayed using the following rule:

$$l = l_0 \times (1 - \frac{i}{N})^{0.9}, \tag{11}$$

where l_0 is the initial learning rate, i is an epoch counter and N is the total number of the epochs. We experimentally set l_0 to 10^{-4} and N to 300.

The labels provided by the BraTS benchmark include the ED, NCR/NET and ET, while the evaluation of segmentation accuracy is performed on three partially overlapping regions: WT (ET + NCR/NET + ED), TC (ET + NCR/NET) and ET, as mentioned in Section 1. In our experiments, we adopt the region-based training strategy, which directly optimizes these three sub-regions instead of individual labels, since its effectiveness has been widely verified in brain tumor segmentation (Isensee et al., 2020). For post-processing, we also adopt a frequently-used approach that the ET is replaced by the NCR/NET when its volume is less than 500 voxels to remove possible false predictions on ET (Isensee et al., 2020; Lyu and Shu, 2020; Zhang et al., 2020a). Two popular objective metrics including the Dice score and the Hausdorff distance (%95) are used to evaluate the segmentation accuracy.

4.2. Parameter analysis

The loss functions in our method contain several trade-off parameters such as α, β, γ, and λ. The principle for determining the values of β and γ has been detailed in Section 3.2. In this subsection, we analyze the effect of the parameters α and λ on the segmentation performance of the proposed method. The parameter α is used to balance the pixel loss and the structural similarity loss, and these two terms should have relatively close values so that both of them can have sufficient contribution. Based on the experimental observations, we set α to 150, 300, 450, 600, and 750 to study its impact. The corresponding results are shown in the first two sub-figures in Figure 6. It can be seen that the proposed method can obtain relatively stable performance when α is set between 150 and 750, and in particular between 300 and 600. Based on these results, we set α to 450 by default in our experiments. The parameter λ controls the balance between the dice loss and the BCE loss in the segmentation model. Similarly, we set λ to 0.1, 0.3, 0.5, 0.7, 0.9 to analyze its effect on the model performance. The corresponding results are given in the last two sub-figures in Figure 6. We can see that the setting of 0.5 can result in the best performance in most cases, so the parameter λ is set to 0.5 by default in our method.

4.3. Ablation study of the proposed method

In this subsection, an ablation study is conducted to evaluate the effectiveness of our PIF-Net and MSFF module in the proposed method. Specifically, the following four models are considered in this study:

- **OURS w/o PIF-Net&MSFF**: Removing the PIF-Net and the MSFF module simultaneously from the proposed brain tumor segmentation framework. In each stage, only the V-Net is remained for segmentation. This is the original baseline for our method.
- **OURS w/o PIF-Net**: Removing the PIF-Net from the proposed brain tumor segmentation framework. The MSFF module is embedded before the V-Net to realize multimodal feature refinement for segmentation in both stages.
- **OURS w/o MSFF**: Removing the MSFF module from the proposed brain tumor segmentation framework. The PIF-Net is used to generate the fused modalities as the additional input of the segmentation model in both stages.
- **OURS**: The complete model proposed in this work.

The evaluation results on the BraTS 2019 and BraTS 2020 benchmarks are listed in Tables 2, 3, respectively. method

TABLE 2 Objective evaluation results for the ablation study of the proposed method on the BraTS 2019 validation sets.

Tumor region	Metrics	OURS w/o PIFnet & MSFF	OURS w/o PIFnet	OURS w/o MSFF	OURS
WT	Dice	0.8635	0.8771	0.8832	**0.8942**
	Hausdorff	7.1211	7.7784	7.1654	**5.3490**
TC	Dice	0.7788	0.8065	0.8045	**0.8142**
	Hausdorff	15.7345	**10.1822**	14.4599	10.8988
ET	Dice	0.7682	0.7698	0.7692	**0.7710**
	Hausdorff	9.1385	**5.3155**	6.4719	5.8548
Average	Dice	0.8035	0.8178	0.8190	**0.8265**
	Hausdorff	10.6647	7.7587	9.3657	**7.3675**

Bold values indicate the best-performing scores on each metric (each row in the tables) among all the four models.

TABLE 3 Objective evaluation results for the ablation study of the proposed method on the BraTS 2020 validation sets.

Tumor region	Metrics	OURS w/o PIFnet & MSFF	OURS w/o PIFnet	OURS w/o MSFF	OURS
WT	Dice	0.8678	0.8725	0.8878	**0.8950**
	Hausdorff	11.5732	9.6274	7.8896	**5.3117**
TC	Dice	0.8025	0.8153	0.8139	**0.8178**
	Hausdorff	11.6728	10.4340	10.9337	**9.4285**
ET	Dice	0.7631	0.7730	0.7678	**0.7745**
	Hausdorff	6.9469	5.9442	7.1674	**4.4715**
Average	Dice	0.8111	0.8203	0.8232	**0.8291**
	Hausdorff	10.0643	8.6685	8.6636	**6.4039**

Bold values indicate the best-performing scores on each metric (each row in the tables) among all the four models.

generally has a better a slightly better performance for BraTS 2020 than performance for BraTS 2020 than BraTS 2019, which is mainly because the BraTS 2020 benchmark contains more training samples in the training set, with additional 34 samples in comparison to the BraTS 2019 benchmark. The comparison between **OURS** and **OURS w/o PIFnet&MSFF** demonstrates that the utilization of our PIF-Net and MSFF module can significantly improve the performance (1.8% to 2.3% in terms of the mean Dice score, and 3.3 to 3.7 in terms of the mean Hausdorff distance) over the baseline model. The comparison between **OURS w/o MSFF** and **OURS w/o PIF-Net&MSFF** (as well as the comparison between **OURS** and **OURS w/o PIF-Net**) verifies the effectiveness of the PIF-Net in improving the segmentation accuracy. The comparison between **OURS w/o PIF-Net** and **OURS w/o PIF-Net&MSFF** (as well as the comparison between **OURS** and **OURS w/o MSFF**) shows that the MSFF module is also beneficial for the segmentation performance. Some segmentation results obtained by **OURS w/o PIF-Net&MSFF**, **OURS w/o PIF-Net**, **OURS w/o MSFF**, and **OURS** are visualized in Figure 7. It can be seen that the complete model can generally obtain more accurate segmentation results than the baseline methods when compared to the ground truth.

An interesting observation we can obtain from Tables 2, 3 are that the improvements achieved by the PIF-Net and the MSFF module have their characteristics on different sub-regions. Specifically, for the WT area, the PIF-Net is more effective in improving the segmentation accuracy than the MSFF module. On the other hand, for the TC and ET areas, the MSFF module is more helpful in comparison to the PIF-Net. This phenomenon can be observed from the comparison between **OURS w/o PIF-Net** and **OURS w/o MSFF**. The results shown in Figure 7 also verify this point. By referring to the ground truth, we can see that **OURS w/o MSFF** generally obtains more accurate results for the ED area (shown in green) than **OURS w/o PIF-Net**, while **OURS w/o PIF-Net** performs better for the NCR/NET and ET areas (shown in red and yellow). We provide an explanation to this observation as follows. The

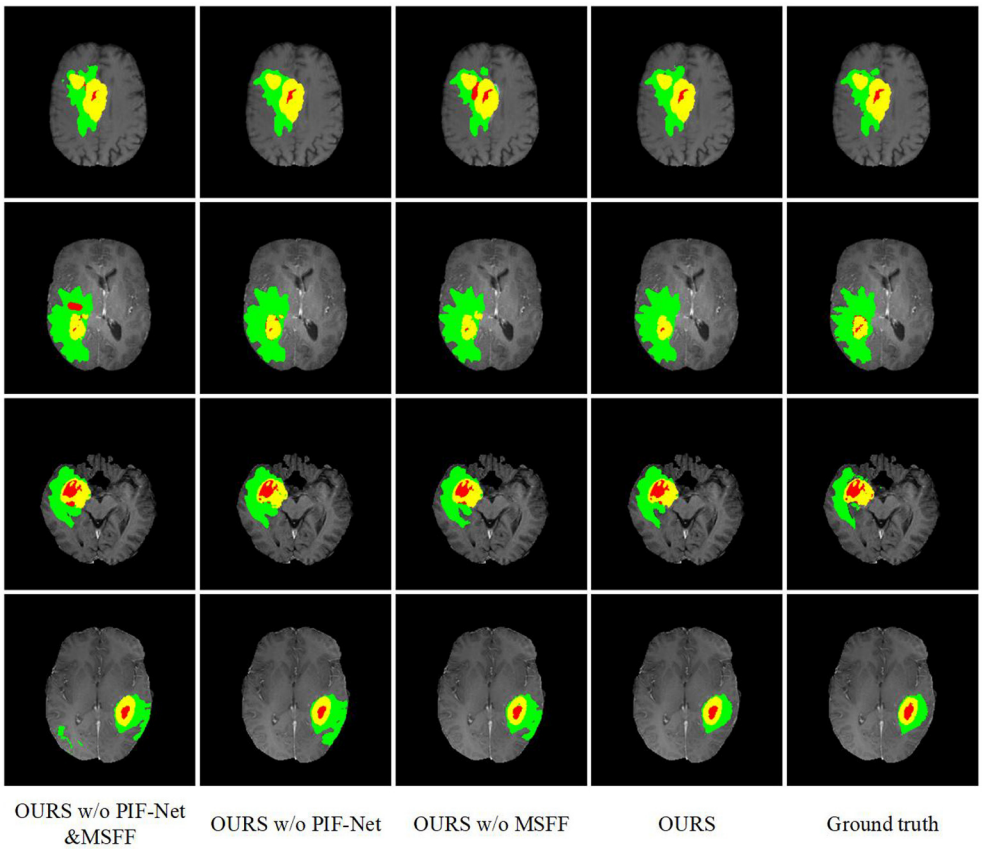

FIGURE 7
Examples of brain tumor segmentation results obtained by di3fferent methods in the ablation study. The green, red, and yellow regions indicate edema (ED), non-enhancing tumor and necrosis (NCR/NET), and enhancing tumor (ET), respectively.

segmentation of WT is mainly based on the ED area that can be effectively captured in the T2 and Flair volumes. The modality characteristics on the ED area in T2 and Flair volumes are generally close, so the requirement of multimodal feature fusion or selection is not very urgent. By contrast, the pixel-level image fusion achieved by the PIF-Net can enrich the input modalities for the segmentation model and this modality enhancement approach can also be viewed as a data augmentation method to some extent, which tends to be relatively more effective for WT segmentation as only two source modalities are used. In comparison to WT, the segmentation of TC and ET is more difficult due to the factors like smaller size, more irregular shape, etc. As a result, more modalities are typically required in TC and ET segmentation. In such a situation, the refinement of multimodal features achieved by the MSFF module is of higher significance. Therefore, the segmentation of TC and ET benefits more from the MSFF module. Nevertheless, it is worth noting that our PIF-Net and MSFF module both improve the segmentation accuracy of all the three sub-regions, just with different extents.

4.4. Comparison with other methods

In this subsection, we compare the proposed method with some existing brain tumor segmentation methods, which are mainly included in the proceedings of BraTS 2019-2021 challenges and generally have good performance. Tables 4, 5 report the evaluation results of different methods on BraTS 2019 and BraTS 2020 validation sets, respectively. For the comparison methods, the results reported in the original publications are adopted since the benchmarks used are exactly the same. In addition, the results obtained by a single model instead of multi-model ensemble are used for the sake of fair comparison. In each case, the best score is indicated in bold and the second best score is underlined. We can observe from Tables 4, 5 that the proposed method achieves very competitive performance among all the methods. For WT and TC regions, the proposed method obtains the highest Dice scores on both BraTS 2019 and BraTS 2020 validation sets. Our method achieves 0.8265 and 0.8291 in terms of the mean Dice score on these two datasets, which are both in the second place among all the methods. It is worth mentioning

TABLE 4 Objective evaluation results of different brain tumor segmentation methods on the BraTS 2019 validation sets.

References	WT		TC		ET		Average	
	Dice	Hausdorff	Dice	Hausdorff	Dice	Hausdorff	Dice	Hausdorff
Xu et al. (2019)	0.8930	6.9640	0.8070	7.6630	0.7590	**4.1930**	0.8197	**6.2733**
Baid et al. (2019)	0.8700	13.3600	0.7700	12.7100	0.7000	6.4500	0.7800	10.8400
González et al. (2019)	0.8882	8.1231	0.7833	7.5618	0.7231	4.9132	0.7982	6.8660
Lorenzo et al. (2019)	0.8904	-	0.7511	-	0.6634	-	0.7683	-
Ahmad et al. (2019)	0.8518	9.0083	0.7576	10.6744	0.6230	8.4683	0.7441	9.3837
Abraham and Khan (2019)	0.8605	-	0.7108	-	0.6323	-	0.7345	-
Bhalerao and Thakur (2019)	0.8527	8.0793	0.7091	9.5708	0.6668	7.2700	0.7429	8.3067
Yan et al. (2019)	0.8600	40.3100	0.7300	10.4000	0.6600	18.5300	0.7500	23.0800
Iantsen et al. (2019)	0.8700	8.3500	0.7900	9.5800	0.6700	7.8200	0.7767	8.5833
Astaraki et al. (2019)	0.8700	5.9000	0.8100	**7.1600**	0.7100	6.0200	0.7967	6.3600
Cao et al. (2021)	0.8938	7.5050	0.7875	9.2600	**0.7849**	6.9250	0.8221	7.8967
Wang et al. (2021)	0.8889	7.5990	0.8141	7.5840	0.7836	5.9080	**0.8289**	7.0303
Valanarasu et al. (2021)	0.8760	8.9420	0.7392	9.8930	0.7321	6.3230	0.7824	8.3860
OURS	**0.8942**	**5.3490**	**0.8142**	10.8988	0.7710	5.8548	0.8265	7.3675

Bold and underlined values indicate the best scores and second best scores on each metric (each column in the tables) among all the methods.

TABLE 5 Objective evaluation results of different brain tumor segmentation methods on the BraTS 2020 validation sets.

References	WT		TC		ET		Average	
	Dice	Hausdorff	Dice	Hausdorff	Dice	Hausdorff	Dice	Hausdorff
Jun et al. (2020)	0.8780	6.3000	0.7790	11.0200	0.7520	30.6500	0.8030	15.9900
Liu et al. (2020a)	0.8823	6.4900	0.8012	**6.6800**	0.7637	21.3900	0.8157	11.5200
Messaoudi et al. (2020)	0.8413	-	0.6804	-	0.6537	-	0.7251	-
Sun et al. (2020)	0.8920	-	0.7880	-	0.7230	-	0.8010	-
Cirillo et al. (2020)	0.8926	6.3900	0.7919	14.0700	0.7504	36.0000	0.8116	18.8200
Pang et al. (2020)	0.8811	18.0901	0.7605	29.0570	0.7538	34.2391	0.7985	27.1287
Sundaresan et al. (2020)	0.8900	**4.4000**	0.7700	15.3000	0.7700	29.4000	0.8100	16.3667
Ballestar and Vilaplana (2020)	0.8300	12.3400	0.7700	13.1100	0.7200	37.4200	0.7733	20.9567
McHugh et al. (2020)	0.8810	6.7200	0.7890	10.2000	0.7120	40.6000	0.7940	19.1733
Ma et al. (2020b)	0.8794	-	0.7731	-	0.7040	-	0.7855	-
Cao et al. (2021)	0.8934	7.855	0.7760	14.5940	**0.7895**	11.0050	0.8196	11.1513
Wang et al. (2021)	0.8900	6.4690	0.8136	10.4680	0.7850	16.7160	**0.8295**	11.2177
Zhang et al. (2021b)	0.8800	6.9500	0.7400	30.1800	0.7000	38.6000	0.7733	25.2433
OURS	**0.8950**	5.3117	**0.8178**	9.4285	0.7745	**4.4715**	0.8291	**6.4039**

Bold and underlined values indicate the best scores and second best scores on each metric (each column in the tables) among all the methods.

that the performance of proposed method may be slightly inferior to some latest state-of-the-art methods. However, the main purpose of this work is to verify the effectiveness of the proposed pixel-level and feature-level image fusion approaches for brain tumor segmentation. The segmentation model and loss function adopted in this work are both plain while popular approaches (i.e., the original V-Net and the BCE-and-Dice-based loss) in 3D medical image segmentation. By introducing some advanced architectures and loss functions, we believe that the segmentation performance can be further improved.

5. Conclusion

In this paper, we mainly introduce pixel-level and feature-level image fusion techniques for MRI-based brain tumor segmentation, aiming to achieve more sufficient and finer utilization of multimodal information. Specifically, we present a CNN-based 3D pixel-level image fusion network named PIF-Net to enrich the input modalities of the segmentation model and design an attention-based feature fusion module named MSFF for multimodal feature refinement. A two-stage

brain tumor segmentation framework is accordingly proposed based on the PIF-Net, the MSFF module and the V-Net. Experimental results on the BraTS 2019 and BraTS 2020 benchmarks show that the proposed components on both pixel-level and feature-level fusion can effectively improve the segmentation accuracy of all the three tumor sub-regions including whole tumor, tumor core and enhancing tumor. The pixel-level image fusion network in this work is trained independently to the segmentation model. Future work may concentrate on integrating image fusion and segmentation into a unified network for better feature learning to further improve the segmentation performance.

Data availability statement

The datasets for this study can be found in the BraTS 2019 dataset available at: https://www.med.upenn.edu/cbica/brats2019/data.html and in the BraTS 2020 dataset available at: https://www.med.upenn.edu/cbica/brats2020/data.html.

Author contributions

YL: conceptualization, methodology, and writing. FM: methodology, experiments, and Writing. YS: methodology and experiments. JC: methodology and review. CL: experiments and review. XC: methodology, review, and supervision. All authors contributed to the work and approved the submission.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grants 62176081, 61922075, 62171176, and 41901350, and in part by the Fundamental Research Funds for the Central Universities under Grants JZ2020HGPA0111 and JZ2021HGPA0061.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Abraham, N., and Khan, N. M. (2019). "Multimodal segmentation with MGF-Net and the focal tversky loss function," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 191–198. doi: 10.1007/978-3-030-46643-5_18

Ahmad, P., Qamar, S., Hashemi, S. R., and Shen, L. (2019). "Hybrid labels for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 158–166. doi: 10.1007/978-3-030-46643-5_15

Astaraki, M., Wang, C., Carrizo, G., Toma-Dasu, I. and Smedby, Ö. (2019). "Multimodal brain tumor segmentation with normal appearance autoencoder," in *International MICCAI Brainlesion Workshop* (Springer), 316–323. doi: 10.1007/978-3-030-46643-5_31

Baid, U., Shah, N. A., and Talbar, S. (2019). "Brain tumor segmentation with cascaded deep convolutional neural network," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 90–98. doi: 10.1007/978-3-030-46643-5_9

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). 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*.

Ballestar, L. M., and Vilaplana, V. (2020). "MRI brain tumor segmentation and uncertainty estimation using 3D-UNet architectures," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 376–390. doi: 10.1007/978-3-030-72084-1_34

Bauer, S., Nolte, L.-P., and Reyes, M. (2011). "Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Berlin, Heidelberg: Springer), 354–361. doi: 10.1007/978-3-642-23626-6_44

Bhalerao, M., and Thakur, S. (2019). "Brain tumor segmentation based on 3D residual U-Net," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 218–225. doi: 10.1007/978-3-030-46643-5_21

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., et al. (2021). Swin-unet: UNet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*. doi: 10.48550/arXiv.2105.05537

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). TransUNet: transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*. doi: 10.48550/arXiv.2102.04306

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-assisted Intervention* (Athens, Greece: Springer), 424–432. doi: 10.1007/978-3-319-46723-8_49

Cirillo, M. D., Abramian, D., and Eklund, A. (2020). "Vox2Vox: 3D-GAN for brain tumour segmentation," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 274–284. doi: 10.1007/978-3-030-72084-1_25

Du, J., Li, W., Lu, K., and Xiao, B. (2016). An overview of multi-modal medical image fusion. *Neurocomputing* 215, 3–20. doi: 10.1016/j.neucom.2015.07.160

Goetz, M., Weber, C., Bloecher, J., Stieltjes, B., Meinzer, H.-P., and Maier-Hein, K. (2014). "Extremely randomized trees based brain tumor segmentation," in *Proceeding of MICCAI BRATS Challenge* (Boston, MA, USA), 6–11.

González, S. R., Sekou, T. B., Hidane, M., and Tauber, C. (2019). "3D automatic brain tumor segmentation using a multiscale input U-Net network," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 113–123. doi: 10.1007/978-3-030-46643-5_11

Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., et al. (2022). Attention mechanisms in computer vision: a survey. *Comp. Visual Media* 8, 331–368. doi: 10.1007/s41095-022-0271-y

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., et al. (2017). Brain tumor segmentation with deep neural networks. *Med. Image Anal.* 35, 18–31. doi: 10.1016/j.media.2016.05.004

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, USA), 7132–7141. doi: 10.1109/CVPR.2018.00745

Iantsen, A., Jaouen, V., Visvikis, D., and Hatt, M. (2019). "Encoder-decoder network for brain tumor segmentation on multi-sequence MRI," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 296–302. doi: 10.1007/978-3-030-46643-5_29

Isensee, F., Jager, P. F., Full, P. M., Vollmuth, P., and Maier-Hein, K. H. (2020). "nnU-Net for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 118–132. doi: 10.1007/978-3-030-72087-2_11

Jun, W., Haoxiang, X., and Wang, Z. (2020). "Brain tumor segmentation using dual-path attention U-Net in 3D MRI images," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 183–193. doi: 10.1007/978-3-030-72084-1_17

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., et al. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* 36, 61–78. doi: 10.1016/j.media.2016.10.004

Kleesiek, J., Biller, A., Urban, G., Kothe, U., Bendszus, M., and Hamprecht, F. (2014). "Ilastik for multi-modal brain tumor segmentation," in *Proceeding of MICCAI BRATS Challenge* (Boston, MA, USA), 12–17.

Li, S., Kang, X., Fang, L., Hu, J., and Yin, H. (2017). Pixel-level image fusion: a survey of the state of the art. *Inform. Fusion* 33, 100–112. doi: 10.1016/j.inffus.2016.05.004

Li, X., Luo, G., and Wang, K. (2019a). "Multi-step cascaded networks for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 163–173. doi: 10.1007/978-3-030-46640-4_16

Li, X., Wang, W., Hu, X., and Yang, J. (2019b). "Selective kernel networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (California, USA), 510–519. doi: 10.1109/CVPR.2019.00060

Liang, X., Hu, P., Zhang, L., Sun, J., and Yin, G. (2019). MCFNet: multi-layer concatenation fusion network for medical images fusion. *IEEE Sensors J.* 19, 7107–7119. doi: 10.1109/JSEN.2019.2913281

Liu, C., Ding, W., Li, L., Zhang, Z., Pei, C., Huang, L., et al. (2020a). "Brain tumor segmentation network using attention-based fusion and spatial relationship constraint," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 219–229. doi: 10.1007/978-3-030-72084-1_20

Liu, J., Li, M., Wang, J., Wu, F., Liu, T., and Pan, Y. (2014). A survey of MRI-based brain tumor segmentation methods. *Tsinghua Sci. Technol.* 19, 578–595. doi: 10.1109/TST.2014.6961028

Liu, Y., Chen, X., Cheng, J., and Peng, H. (2017). "A medical image fusion method based on convolutional neural networks," in *2017 20th International Conference on Information Fusion (Fusion)* (Xian, China), 1070–1076. doi: 10.23919/ICIF.2017.8009769

Liu, Y., Chen, X., Liu, A., Ward, R. K., and Wang, Z. J. (2021). Recent advnaces in sparse representation based medical image fusion. *IEEE Instrument. Meas. Mag.* 24, 45–53. doi: 10.1109/MIM.2021.9400960

Liu, Y., Chen, X., Wang, Z., Wang, Z., Ward, R., and Wang, X. (2018). Deep learning for pixel-level image fusion: recent advances and future prospects. *Inform. Fusion* 42, 158–173. doi: 10.1016/j.inffus.2017.10.007

Liu, Y., Chen, X., Ward, R. K., and Wang, Z. J. (2016). Image fusion with convolutional sparse representation. *IEEE Signal Process. Lett.* 23, 1882–1886. doi: 10.1109/LSP.2016.2618776

Liu, Y., Chen, X., Ward, R. K., and Wang, Z. J. (2019). Medical image fusion via convolutional sparsity based morphological component analysis. *IEEE Signal Process. Lett.* 26, 485–489. doi: 10.1109/LSP.2019.2895749

Liu, Y., Liu, S., and Wang, Z. (2015). A general framework for image fusion based on multi-scale transform and sparse representation. *Inform. Fusion* 24, 147–164. doi: 10.1016/j.inffus.2014.09.004

Liu, Y., Shi, Y., Mu, F., Cheng, J., Li, C., and Chen, X. (2022). Multimodal mri volumetric data fusion with convolutional neural networks. *IEEE Trans, Instrument. Meas.* 71, 4006015. doi: 10.1109/TIM.2022.3184360

Liu, Y., Wang, L., Cheng, J., Li, C., and Chen, X. (2020b). Multi-focus image fusion: a survey of the state of the art. *Inform. Fusion* 64, 71–91. doi: 10.1016/j.inffus.2020.06.013

Lorenzo, P. R., Marcinkiewicz, M., and Nalepa, J. (2019). "Multi-modal U-Nets with boundary loss and pre-training for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 135–147. doi: 10.1007/978-3-030-46643-5_13

Lyu, C., and Shu, H. (2020). "A two-stage cascade model with variational autoencoders and attention gates for MRI brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 435–447. doi: 10.1007/978-3-030-72084-1_39

Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., and Ma, Y. (2022). SwinFusion: cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA. Automat. Sin.* 9, 1200–1217. doi: 10.1109/JAS.2022.105686

Ma, J., Xu, H., Jiang, J., Mei, X., and Zhang, X.-P. (2020a). DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans. Image Process.* 29, 4980–4995. doi: 10.1109/TIP.2020.2977573

Ma, J., and Yang, X. (2018). "Automatic brain tumor segmentation by exploring the multi-modality complementary information and cascaded 3D lightweight CNNs," in *International MICCAI Brainlesion Workshop* (Granada, Spain: Springer), 25–36. doi: 10.1007/978-3-030-11726-9_3

Ma, S., Zhang, Z., Ding, J., Li, X., Tang, J., and Guo, F. (2020b). "A deep supervision CNN network for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 158–167. doi: 10.1007/978-3-030-72087-2_14

McHugh, H., Talou, G. M., and Wang, A. (2020). "2d Dense-UNet: a clinically valid approach to automated glioma segmentation," in *International MICCAI Brainlesion Workshop* (Springer), 69–80. doi: 10.1007/978-3-030-72087-2_7

Meier, R., Bauer, S., Slotboom, J., Wiest, R., and Reyes, M. (2014). "Patient-specific semi-supervised learning for postoperative brain tumor segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Boston, MA, USA: Springer), 714–721. doi: 10.1007/978-3-319-10404-1_89

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Messaoudi, H., Belaid, A., Allaoui, M. L., Zetout, A., Allili, M. S., Tliba, S., et al. (2020). "Efficient embedding network for 3D brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 252–262. doi: 10.1007/978-3-030-72084-1_23

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-Net: fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)* (Stanford, CA, USA: IEEE), 565–571. doi: 10.1109/3DV.2016.79

Myronenko, A. (2018). "3D MRI brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop* (Granada, Spain: Springer), 311–320. doi: 10.1007/978-3-030-11726-9_28

Pang, E., Shi, W., Li, X., and Wu, Q. (2020). "Glioma segmentation using encoder-decoder network and survival prediction based on cox analysis," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 318–326. doi: 10.1007/978-3-030-72084-1_29

Pinto, A., Pereira, S., Correia, H., Oliveira, J., Rasteiro, D. M., and Silva, C. A. (2015). "Brain tumour segmentation based on extremely randomized forest with high-level features," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Milan, Italy: IEEE), 3037–3040. doi: 10.1109/EMBC.2015.7319032

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention (MICCAI)* (Munich, Germany: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Subbanna, N. K., Precup, D., Collins, D. L., and Arbel, T. (2013). "Hierarchical probabilistic Gabor and MRF segmentation of brain tumours in MRI volumes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Nagoya, Japan), 751–758. doi: 10.1007/978-3-642-40811-3_94

Sun, J., Peng, Y., Li, D., and Guo, Y. (2020). "Segmentation of the multimodal brain tumor images used Res-U-Net," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 263–273. doi: 10.1007/978-3-030-72084-1_24

Sundaresan, V., Griffanti, L., and Jenkinson, M. (2020). "Brain tumour segmentation using a triplanar ensemble of U-Nets on MR images," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 340–353. doi: 10.1007/978-3-030-72084-1_31

Tang, W., Liu, Y., Cheng, J., Li, C., and Chen, X. (2021). Green fluorescent protein and phase contrast image fusion via detail preserving cross network. *IEEE Trans. Comput. Imaging* 7, 584–597. doi: 10.1109/TCI.2021.3083965

Valanarasu, J. M. J., Sindagi, V. A., Hacihaliloglu, I., and Patel, V. M. (2021). Kiu-Net: overcomplete convolutional architectures for biomedical image and volumetric segmentation. *IEEE Trans. Med. Imaging* 41, 965–976. doi: 10.1109/TMI.2021.3130469

Wang, G., Li, W., Ourselin, S., and Vercauteren, T. (2017). "Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks," in *International MICCAI Brainlesion Workshop* (Quebec City, QC, Canada: Springer), 178–190. doi: 10.1007/978-3-319-75238-9_16

Wang, L., Li, B., and Tian, L. (2014). Multimodal medical volumetric data fusion using 3-D discrete shearlet transform and global-to-local rule. *IEEE Trans. Biomed. Eng.* 61, 197–206. doi: 10.1109/TBME.2013.2279301

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., and Li, J. (2021). "TransBTS: Multimodal brain tumor segmentation using transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Strasbourg, France: Springer), 109–119. doi: 10.1007/978-3-030-87193-2_11

Xu, H., and Ma, J. (2021). EMFusion: an unsupervised enhanced medical image fusion network. *Inform. Fusion* 76, 177–186. doi: 10.1016/j.inffus.2021.06.001

Xu, H., Ma, J., Jiang, J., Guo, X., and Ling, H. (2022). U2Fusion: a unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 502–518. doi: 10.1109/TPAMI.2020.3012548

Xu, X., Zhao, W., and Zhao, J. (2019). "Brain tumor segmentation using attention-based network in 3D MRI images," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 3–13. doi: 10.1007/978-3-030-46643-5_1

Yan, K., Sun, Q., Li, L., and Li, Z. (2019). "3D Deep residual encoder-decoder CNNS with squeeze-and-excitation for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 234–243. doi: 10.1007/978-3-030-46643-5_23

Yang, Y., Que, Y., Huang, S., and Lin, P. (2016). Multimodal sensor medical image fusion based on type-2 fuzzy logic in nsct domain. *IEEE Sensors J.* 16, 3735–3745. doi: 10.1109/JSEN.2016.2533864

Yin, H. (2018). Tensor sparse representation for 3-D medical image fusion using weighted average rule. *IEEE Trans. Biomed. Eng.* 65, 2622–2633. doi: 10.1109/TBME.2018.2811243

Yin, M., Liu, X., Liu, Y., and Chen, X. (2019). Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain. *IEEE Trans. Instrument. Meas.* 68, 49–64. doi: 10.1109/TIM.2018.2838778

Zhang, D., Huang, G., Zhang, Q., Han, J., Han, J., Wang, Y., et al. (2020a). Exploring task structure for brain tumor segmentation from multi-modality MR images. *IEEE Trans. Image Process.* 29, 9032–9043. doi: 10.1109/TIP.2020.3023609

Zhang, H., Xu, H., Tian, X., Jiang, J., and Ma, J. (2021a). Image fusion meets deep learning: a survey and perspective. *Inform. Fusion* 76, 323–336. doi: 10.1016/j.inffus.2021.06.008

Zhang, Q., Liu, Y., Blum, R., Han, J., and Tao, D. (2018). Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: a review. *Inform. Fusion* 40, 57–75. doi: 10.1016/j.inffus.2017.05.006

Zhang, W., Yang, G., Huang, H., Yang, W., Xu, X., Liu, Y., et al. (2021b). ME-Net: multi-encoder net framework for brain tumor segmentation. *Int. J. Imaging Syst. Technol.* 31, 1834–1848. doi: 10.1002/ima.22571

Zhang, Y., He, N., Yang, J., Li, Y., Wei, D., Huang, Y., et al. (2022). mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. *arXiv preprint arXiv:2206.02425*. doi: 10.48550/arXiv.2206.02425

Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., and Zhang, L. (2020b). IFCNN: a general image fusion framework based on convolutional neural network. *Inform. Fusion* 54, 99–118. doi: 10.1016/j.inffus.2019.07.011

Zhao, X., Wu, Y., Song, G., Li, Z., Zhang, Y., and Fan, Y. (2018). A deep learning model integrating FCNNs and CRFs for brain tumor segmentation. *Med. Image Anal.* 43, 98–111. doi: 10.1016/j.media.2017.10.002

Zhou, C., Ding, C., Wang, X., Lu, Z., and Tao, D. (2020). One-pass multi-task networks with cross-task guided attention for brain tumor segmentation. *IEEE Trans. Image Process.* 29, 4516–4529. doi: 10.1109/TIP.2020.2973510

Zhu, Z., Yin, H., Chai, Y., Li, Y., and Qi, G. (2018). A novel multi-modality image fusion method based on image decomposition and sparse representation. *Inform. Sci.* 432, 516–529. doi: 10.1016/j.ins.2017.09.010