Open access • Journal Article • DOI:10.1109/TIP.2020.3023609

Exploring Task Structure for Brain Tumor Segmentation From Multi-Modality MR Images — Source link

Dingwen Zhang, Guohai Huang, Qiang Zhang, Jungong Han ...+3 more authors

Institutions: Xidian University, Aberystwyth University, Northwestern Polytechnical University, Peking University

Published on: 17 Sep 2020 - IEEE Transactions on Image Processing (IEEE)

Topics: Image segmentation and Segmentation

Exploring Task Structure for Brain Tumor Segmentation from Multi-modality MR Images

Zhang, Dingwen; Huang, Guohai; Zhang, Qiang; Han, Jungong; Han, Junwei; Wang, Yizhou

Published in:
IEEE Transactions on Image Processing

DOI:
10.1109/TIP.2020.3023609

Publication date:
2020

Citation for published version (APA):
Zhang, D., Huang, G., Zhang, Q., Han, J., Han, J., & Wang, Y. (2020). Exploring Task Structure for Brain Tumor Segmentation from Multi-modality MR Images. *IEEE Transactions on Image Processing*, *29*, 9032 - 9043. https://doi.org/10.1109/TIP.2020.3023609

Exploring Task Structure for Brain Tumor Segmentation from Multi-modality MR Images

Dingwen Zhang, Guohai Huang, Qiang Zhang, Jungong Han, Junwei Han *Senior Member*, *IEEE*, Yizhou Wang, Yizhou Yu, *Fellow*, *IEEE*

***Abstract*—Brain tumor segmentation, which aims at segmenting the whole tumor area, enhancing tumor core area, and tumor core area from each input multi-modality bio-imaging data, has received considerable attention from both academia and industry. However, the existing approaches usually treat this problem as a common semantic segmentation task without taking into account the underlying rules in clinical practice. In reality, physicians tend to discover different tumor areas by weighing different modality volume data. Also, they initially segment the most distinct tumor area, and then gradually search around to find the other two. We refer to the first property as the task-modality structure while the second property as the task-task structure, based on which we propose a novel task-structured brain tumor segmentation network (TSBTS net). Specifically, to explore the task-modality structure, we design a modality-aware feature embedding mechanism to infer the important weights of the modality data during network learning. To explore the task-task structure, we formulate the prediction of the different tumor areas as conditional dependency sub-tasks and encode such dependency in the network stream. Experiments on BraTS benchmarks show that the proposed method achieves superior performance in segmenting the desired brain tumor areas while requiring relatively lower computational costs, compared to other state-of-the-art methods and baseline models.**

I. Introduction

Brain tumor segmentation aims at automatically segmenting tumor areas from multi-modality Magnetic Resonance (MR) sequences that are imaged by the advanced medical imaging equipment. Through segmenting brain tumors, the volume, shape, and localization of brain tumor areas (including the whole tumor areas, enhancing tumor core areas, and tumor core areas) can be provided, which play crucial roles in brain tumor diagnosis and monitoring[1]. However, segmenting brain tumors from noisy medical images is never an easy task and many research efforts have been devoted to this area, which generally follow two main pathways. On one hand, the existing approaches consider the multi-modality brain tumor segmentation task as a common semantic segmentation problem and build their models based on the network architectures for semantic segmentation [1], [2], [3]. On the other hand, several existing approaches further extend the 2D convolutional neural network (CNN) architectures that are commonly used in semantic segmentation into the 3D CNN architectures [4], [5] to fit the data structure of the investigated multi-modality MR volumes.

However, intending to replicate semantic segmentation methods for RGB images, the existing approaches for brain tumor segmentation seem to rely too much on the CNN architectures, while ignoring the underlying rules for identifying brain tumor areas in clinical practice. Thus, the performance of these approaches is still not satisfactory. In fact, brain disease physicians usually discover different tumor areas by weighing different modality volume data because they know that different modality data may reflect different pathological features. This reveals the underlying **task-modality structure** in brain tumor segmentation, and indicates the relationship between each modality data and the interested tumor area. On the other hand, **physicians in brain disease neither seek the three tumor areas simultaneously nor do they treat each modality equally to find a certain tumor area**. To our best knowledge, this is because physicians have the task structure prior in mind: On one hand, they know that the three tumor areas are mutually included rather than being located independently. Thus, they find these tumor areas by first localizing the most distinct one and then searching around to find the others. This implies the underlying **task-task structure** in brain tumor segmentation if we treat the segmentation of each type of tumor area as a single sub-task. These properties are illustrated in Fig. 1.

Manuscript received xx; revised xx; accepted xx. Date of publication XX; date of current version XX. This paper was recommended by XX. (Corresponding authors: Qiang Zhang and Jungong Han.)

Dingwen Zhang, Guohai Huang, and Qiang Zhang are with School of Mechano-Electronic Engineering, Xidian University, Xian, China. Email: zhangdingwen2006yyy@gmail.com.

Jungong Han is with Computer Science Department, Aberystwyth University, Ceredigion, SY23 3FL, UK. jungonghan77@gmail.com

Junwei Han is with School of Automation, Northwestern Polytechnical University, Xi'an, China. E-mail: junweihan2010@gmail.com.

Yizhou Wang is with School of Electronics Engineering and Computer Science, Peking University, Beijing, China. E-mail: Yizhou.Wang@pku.edu.cn.

Yizhou Yu is with Deepwise AI Lab, Beijing, China. E-mail: yizhouy@acm.org.

[1]The brain tumors studied in this work are the Low-grade gliomas (LGG) and High-grade gliomas (HGG) as 1) gliomas are the most common primary brain malignancies and automatic gliomas segmentation algorithms could alleviate huge human labor in brain tumor diagnosis and monitoring; 2) segmenting gliomas in multimodal MRI scans is one of the most challenging tasks in medical image analysis due to the highly heterogeneous appearance and shape; and 3) compared with the other gliomas, i.e., the Biologically benign gliomas, LGG and HGG will have a prognosis resulting in eventual death.

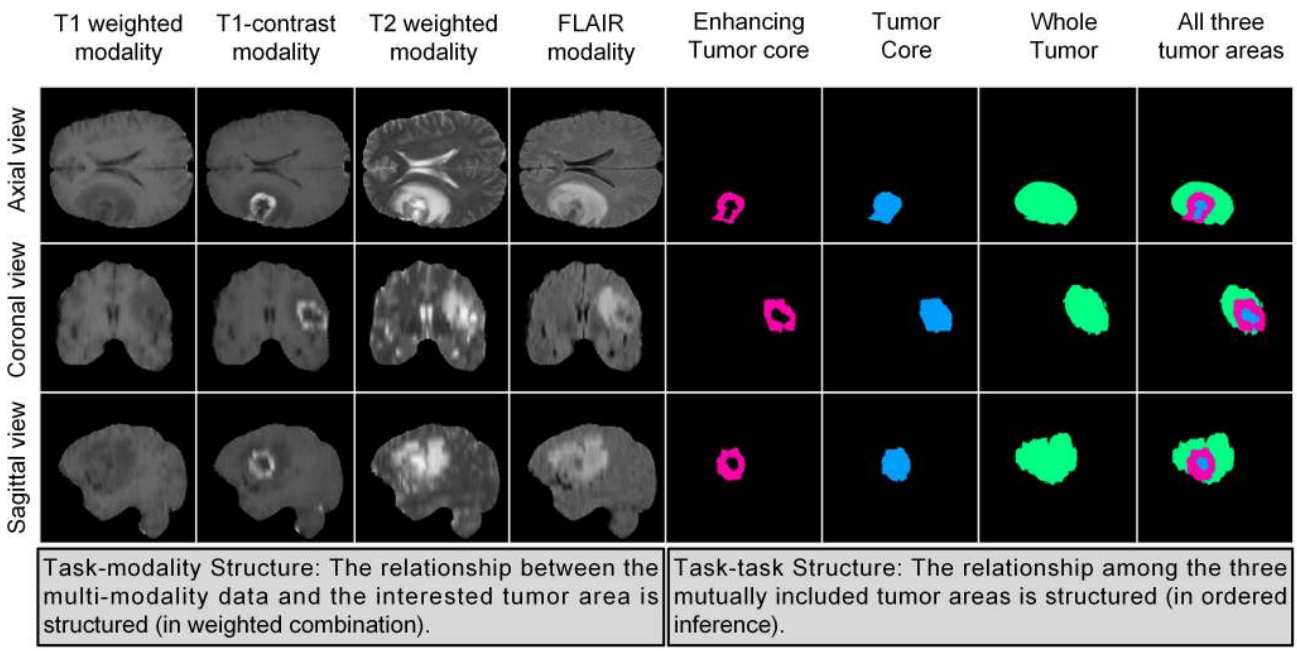

Fig. 1: Illustration of the multi-modality brain tumor segmentation task. On the left, we show MR scans in different modalities and from different views (imagining the entire multi-modality input data is a 3D volume with four modality channels). On the right, we show the ground-truth of the targeted tumor areas. At the bottom of this figure, we reveal two valuable prior knowledge from the clinical practice, which are also the key insights for establishing our proposed TSBTS net. Notice that, to show examples more clearly, we reduce the black background areas for each slice.

Inspired by the aforementioned clinical practices, we propose a novel task-structured brain tumor segmentation network (TSBTS net), which is designed to explore the task structures, including both the task-modality structure and task-task structure, to mimic the physicians' expertise. On the one hand, we model the relationship between the multi-modality data and the targeted tumor areas in a weighted combination structure, where the weights indicating the importance of each modality data for segmenting a particular tumor area are formulated in a modality-aware feature embedding (MAFE) module of the proposed network model. On the other hand, we model the relationship among the sub-tasks on segmenting the three mutually included tumor areas in an ordered inference structure, where the segmentation processes of different tumor areas are formulated as the conditionally dependent sub-tasks and such dependency is encoded in the proposed network stream.

The concrete framework is shown in Fig. 2. As can be seen, we introduce three MAFE modules to infer the importance weights for each modality data and get the weighted features for segmenting the tumor areas. It is worth mentioning that instead of only using a single MAFE module, we use three MAFE modules, each of which corresponds to segmenting a certain type of tumor areas. This design compiles the task-modality structure of the brain tumor segmentation well as it enables our TSBTS net to segment different tumor areas by weighing different modality data. Besides, we can also observe that the mainstream of the proposed network is a feed-forward network, which mainly consists of three inferring modules. Thus, unlike the conventional architectures that simultaneously predict the segmentation maps of all types of brain tumor areas at the end of the network, we instead predict these brain tumor areas separately in different learning modules—-predicting the enhancing tumor core (ET) area from Inferring Module I, the tumor core (TC) area from Inferring Module II, and the whole tumor (WT) area from Inferring Module III, respectively. By mimicking the process to discover and segment the three mutually included areas from the most distinct one to the other surrounding ones, this design can properly encode the task-task structure in brain tumor segmentation.

To sum up, this work mainly contains the following three-fold contributions:

- Inspired by the clinical practice, we reveal the insight of the task structures (including the task-modality structure and the task-task structure) for multi-modality brain tumor segmentation and build the novel task-structured brain tumor segmentation network (TSBTS net).
- We model the task-modality structure as a weighted combination structure and the task-task structure as an ordered inference structure. The former is achieved by modality-aware feature embedding for each particular sub-task while the latter is achieved by formulating the conditional dependency between sub-tasks.
- Comprehensive experiments on BraTS 2017 and 2018 datasets have been conducted to demonstrate that our proposed approach outperforms the baseline models

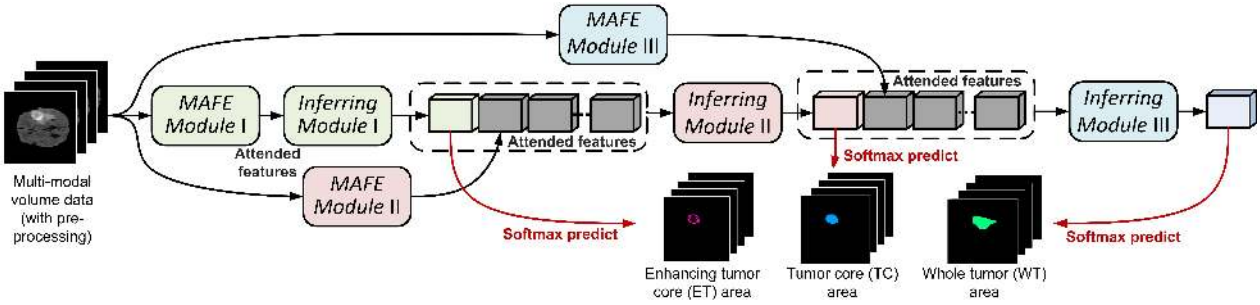

Fig. 2: Illustration of the whole network architecture of the proposed TSBTS net, which contains three modality-aware feature embedding (MAFE) modules and three inferring modules. The modules in green, red, and blue are used to segment the enhancing tumor core, tumor core, and whole tumor areas, respectively. In the TSBTS net, the task-modality structure is modeled as a weighted combination structure, where the MAFE mechanism is adopted to infer the importance weights and obtain the weighted features. In addition, the task-task structure is modeled as an ordered inference structure, where the network mimics the process to discover and segment the three mutually contained areas from the most distinct area to its surrounding areas. Notice that the dashed blocks are not network layers but the input data of each inferring module.

and achieves the state-of-the-art performance.

II. Previous Works

For segmenting brain tumors automatically, researchers in the fields of computer vision and machine learning have made great efforts in the past few decades. In early ages, researchers addressed this problem by mainly using hand-crafted features (such as the context feature [6], gradient feature [7], symmetry feature [8], and physical feature [9]) and shallow learning models (such as Conditional Random Field [10], Support Vector Machines [11], and Random Forests [12]). For example, Tustison et al. [9] used the Random Forests to build a two-stage brain tumor segmentation framework, where the output of the first classifier was used to improve the second stage of segmentation. Geremia et al. [13] developed a hierarchical semination framework based on a Spatially Adaptive Random Forests model. Meier et al. [14] proposed a semi-supervised learning approach to train a subject-specific classifier for post-operative brain tumor segmentation.

More recently, with the rapid development of the deep learning technique [17], deep neural networks (DNNs) with different network architectures have been established to address the brain tumor segmentation problem. Compared with the conventional approaches, the biggest advantage of the DNN-based brain tumor segmentation approach is that useful features could be learned, alongside the targeted segmentation task, in a data-driven manner.

The DNN-based brain tumor segmentation methods can be divided into two categories. The first category is the 2D CNN-based methods. These methods split the multi-modality 3D volume data into 2D patches or slices and use CNNs with 2D convolution operation to process each 2D patch or slice. For example, Shaikh et al. [18] proposed a 100-layer Tiramisu architecture, which integrates a densely connected fully convolutional neural network (FCNN) followed by a Dense Conditional Random Field (DCRF), to segment brain tumors from multi-modal MR slices. Similarly, Islam and Ren [19] extracted the hypercolumn features from FCNN to predict the segmentation masks of each MR slice. Lopez et al. [20] introduced the dilation operation into the deep network for 2D slice-based brain segmentation. They also studied the class imbalance issue for segmenting different tumor areas.

The other category is the 3D CNN-based methods. These methods use CNNs with 3D convolution operation to process the whole MR 3D volume data or the extracted 3D patches. One of the most representative 3D CNN-based brain tumor segmentation approaches is DeepMedic [21], where a dual pathway 3D CNN with 11 layers is proposed to perform on the local image patches for brain tumor segmentation. The network processes the input image patch at multiple scales and the obtained result is further refined by a fully connected Conditional Random Field (CRF). Besides, Li et al. [22] proposed a compact end-to-end 3D CNN model, where high-resolution multi-scale features are maintained with dilated convolutions and residual connections. Castillo et al. [23] developed a volumetric multi-modality neural network organized in three parallel pathways with different input resolutions to predict the labels of each 3D column patch.

Among the existing works, [24] and [25] are the most relevant ones. Specifically, in [24], Mohseni et al. proposed an auto-context convolutional neural network (Auto-Net) for extracting brain areas from the magnetic resonance images. It can be seen as a hierarchy of classifiers, where the objective of each classifier is identical. However, in our approach, the objective of each classifier is for completing different (sub-)tasks and different task-modality relationships are learned for each specific sub-task. In other words, [24] formulates the segmentation problem as a structured prediction problem, which considers the spatial relationship among multiple pixels, i.e., global or local neighborhood interactions. In contrast, our work considers the relationship among multiple related sub-tasks instead, and the prediction for each sub-task is formulated as a structured prediction

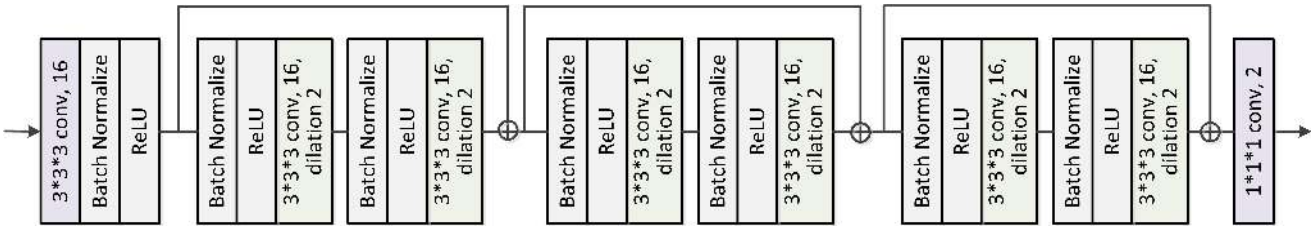

Fig. 3: Illustration of the network architecture of Inferring Module I. The input of the network block is the weighted feature maps obtained from the MAFE Module I, while the output of the network block is the prediction of the ET area. All the convolutional layers perform the 3D voxel-level convolution. The second parameter shown in each convolutional layer is the number of the convolutional kernels. Inspired by [15], [16], we connect convolution modules with post-activation (i.e., Conv-BN-ReLU) and those with pre-activation (i.e., BN-ReLU-Conv) in our network. Notice that the difference between these inferring modules is that the kernel number and the dilation rate of the dilated convolutional layer in the Inferring Module II are set to 32 and 4, respectively. Accordingly, the kernel number and the dilation rate of the dilated convolutional layer in the Inferring Module III are set to 64 and 8.

problem. The network model proposed by [25] also consists of a series of classifier branches which are used to segment the Edema area, Enhancing core area, Necrosis, and Non-enhancing core area in order. However, by concatenating the data from multiple modalities and treating them with equal importance, [25] ignores the exploration of the task-modality structure for this task. Besides, the t-test loss proposed in this work is also new, compared to [24] and [25]. It is worth mentioning that although the main target of our algorithm is to segment the specific types of gliomas, it could succeed in segmenting other tumor types or different tumor grades (with more simple or more complex structures) as long as the data distributions do not change too much from training scenarios to test scenarios.

III. Methodology

The existing works usually formulate the brain tumor segmentation as a structured prediction problem, where the "structure" mainly refers to the context structure, i.e., global or local neighborhood interactions. In this paper, alternatively, we interpret it as a task structured prediction problem based on the valuable domain knowledge from clinical practice. Given the training data $\{\mathbf{X}_m, \mathbf{Y}_m\}, m \in \{1, 2, \cdots, M\}$, where $\mathbf{X}_m = \{\mathbf{X}_m^{T1}, \mathbf{X}_m^{T1c}, \mathbf{X}_m^{T2}, \mathbf{X}_m^{FLAIR}\}$ is the multi-modality volume data while $\mathbf{Y}_m = \{\mathbf{Y}_m^{ET}, \mathbf{Y}_m^{TC}, \mathbf{Y}_m^{WT}\}$ denotes the annotated ground-truth tumor areas, the goal of the proposed TSBTS net is to learn to predict the segmentation masks of the tumor areas $\hat{\mathbf{Y}}_m = \{\hat{\mathbf{Y}}_m^{ET}, \hat{\mathbf{Y}}_m^{TC}, \hat{\mathbf{Y}}_m^{WT}\}$ from each input $\mathbf{X}_m$. Considering no task structure, the existing methods, such as [19], [20], [23], would directly predict labels for multiple (sub-)tasks simultaneously and use unweighted modality features as input. In contrast, our proposed TSBTS net explores the two-fold important task structures, i.e., the task-task structure and task-modality structure, to implement the multi-modality brain tumor segmentation task, which formulates the prediction of multiple (sub-)tasks in a conditional dependency fashion and introduces the modality-aware feature weighting mechanism for each specific (sub-)task.

A. The Network Architecture

As shown in Fig. 2, the basic architecture of our proposed TSBTS net is a feed-forward network, mainly consisting of the Inferring Module I, Inferring Module II, and Inferring Module III. However, different from conventional architectures, we do not simultaneously predict the segmentation maps of all types of brain tumor areas at the end of the network. Instead, we encode the conditional dependency of the sub-tasks into the network stream, which predicts the ET, TC, and WT areas from the three inferring modules in order. Specifically, we first predict the ET area from the Inferring Module I. Then we concatenate the inferred ET area with the pre-computed features as the input of the Inferring Module II and then use the Inferring Module II to predict the TC area. As the TC area is actually around the ET area, the Inferring Module II can leverage the predicted ET area to better infer the TC area. Similarly, after predicting the TC area, we concatenate the predicted TC area with the pre-computed features as the input of the Inferring Module III and use the Inferring Module III to predict the surrounding WT area. The concrete network architecture for each inferring module is shown in Fig. 3, which contains a $3\times3\times3$-voxel convolutional layer and six dilated $3 \times 3 \times 3$-voxel convolutional layers with residual connections.

Besides, before each inferring module, we use a MAFE module to explore the task-modality structure for segmenting brain tumor areas. This is implemented by extracting modality-aware features that maximize the informative patterns for characterizing the corresponding tumor areas. As the volume data of each modality contains different amounts of information when segmenting different types of tumor areas, we use three MAFE modules for extracting modality-aware features corresponding to each type of tumor area. The architecture of the MAFE module is shown in Fig. 4. Specifically, the original $h \times w \times l$-dimensional input multi-modality volume data (h and w indicate the height and width of the volume data, respectively, $l = d \times c$ indicates the depth of the volume data, d and c refer to the number of slices and modalities, respectively) first

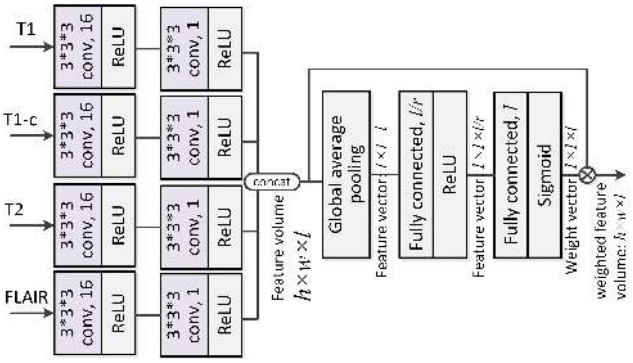

Fig. 4: Illustration of the architecture of the MAFE module. The input of the network block is the original multi-modality volume data, while the output of the network block is the weighted feature volume considering the importance of each modality conditioned on each specific MR slice. Notice that MAFE Module I, II, III share the same architecture. They are placed at different locations of the network to explore the importance of each modality data for segmenting different types of tumor areas.

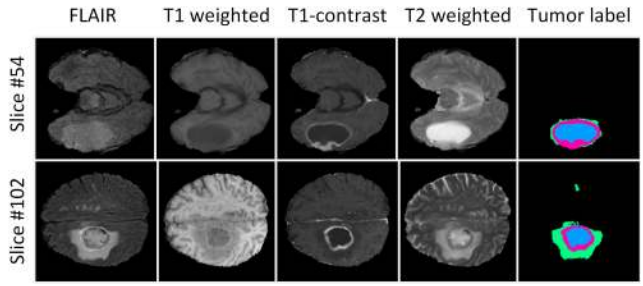

Fig. 5: Examples to explain why we need to learn importance weights conditioned on the specific MR slice location. The examples in the first row are from Slice #54, from which we can observe that the T2 modality is more important for segmenting the WT area. However, from the examples of Slice #102 (the second row), we observe that the FLAIR modality is more important for segmenting the WT area. When segmenting the TC area, the T1 modality appears to be more important in Slice #54 while less important in Slice #102.

undergoes two modality-wise 3D/2D convolutional layers with ReLU to learn features within each modality data. Then, the global average pooling [26] is adopted to embed the feature maps into a l-dimensional vector. After two fully connected layers, the l-dimensional vector is obtained to infer the importance weights for each modality conditioned on each specific MR slice. Finally, the inferred importance weights are multiplied with the previously obtained modality features to generate the weighted feature volume, which is the output of a MAFE module.

Here we explore the importance weights for each modality conditioned on the specific MR slice as we observe that the same modality may have different imaging quality at different slices, while for the same slice, the importance of different modalities is also different (see Fig. 5). This phenomenon can also be observed in Fig. 1, where the imaging quality of the T2 modality data in the top row is better than those in the bottom rows. Consequently, we infer l-dimensional, instead of c-dimensional, importance weights in our MAFE module. After obtaining the $h \times w \times l$-dimensional output feature, we use a split operation to convert it to $h \times w \times d \times c$, indicating c slices with $h \times w \times d$ feature maps for each. Thus, 3D convolution can be used in the network layers of the following inferring module.

From Fig. 3 and 4, it can be observed that by considering the tradeoff between computational complexity and the size of the reception filed, we mainly set the size of the convolutional kernels in the inferring module and MAFE module as 3x3, which is a common choice of most existing works in network design.

B. Task-Task Structure

As we know, the goal of brain tumor segmentation is to segment the ET, TC, and WT tumor areas from the input MR scans. If we treat the segmentation of each type of tumor area as a sub-task, then we can follow the clinical practice to find the strong relationship among these sub-tasks: When segmenting the brain tumor areas, the ET area is always included in the TC area, while the TC area is always included in the WT area (see the right part in Fig. 1). Besides, the ET area tends to be the most attractive tumor area as it is quite distinctive in the T1c modality (see the second column in the left part and the first column in the right part of Fig. 1). Thus, physicians tend to find it in the first instance. Keeping such a task-task structure in mind, we segment the TC area by expanding the ET area and segment the WT area by expanding the TC area.

In light of the above analysis, we model the relationship among the segmentation sub-tasks of the three mutually included tumor areas in an ordered inference structure, where the segmentation processes of the three different tumor areas are formulated as conditional dependency sub-tasks. To be specific, conventional approaches use a deep neural network model $f(\cdot)$ with parameter $\mathbf{W}$ to learn to predict the three types of tumor areas simultaneously:

$$[P(\hat{y}_{m,i}^{ET}=1), P(\hat{y}_{m,i}^{TC}=1), P(\hat{y}_{m,i}^{WT}=1)]^T = f(\mathbf{X}_m; \mathbf{W}); \tag{1}$$

where $\hat{y}_{m,i}^{ET}, \hat{y}_{m,i}^{TC}, \hat{y}_{m,i}^{WT}$ indicate the i-th voxel of the segmentation masks $\hat{\mathbf{Y}}_m^{ET}$, $\hat{\mathbf{Y}}_m^{TC}$, $\hat{\mathbf{Y}}_m^{WT}$, respectively. Different from the conventional approaches, we formulate the conditional dependency of the three related sub-tasks as:

$$\begin{cases} P(\hat{y}_{m,i}^{ET}=1|\mathbf{X}_m) = f_I(\mathbf{X}_m; \mathbf{W}_I); \\ P(\hat{y}_{m,i}^{TC}=1|\hat{\mathbf{Y}}_m^{ET}, \mathbf{X}_m) = f_{II}(\hat{\mathbf{Y}}_m^{ET}, \mathbf{X}_m; \mathbf{W}_{II}); \\ P(\hat{y}_{m,i}^{WT}=1|\hat{\mathbf{Y}}_m^{TC}, \mathbf{X}_m) = f_{III}(\hat{\mathbf{Y}}_m^{TC}, \mathbf{X}_m; \mathbf{W}_{III}), \end{cases} \tag{2}$$

where $\mathbf{W}_I, \mathbf{W}_{II}, \mathbf{W}_{III}$ are the network parameters of the three successive inferring modules $f_I(\cdot), f_{II}(\cdot), f_{III}(\cdot)$, respectively.

C. Task-Modality Structure

From the clinical practice and the example shown in Fig. 1, we can observe that different modalities are of different importance for segmenting a certain type of tumor area, and even the same modality is of different importance for segmenting different types of tumor areas. Thus, when we segment a certain type of tumor area, it is more reasonable to weigh differently on each modality data, rather than treating all of them equally. Such a task-modality structure can be easily modeled as the weighted combination structure and we formulate the learning of these importance weights under the modality-aware feature embedding mechanism:

$$\begin{aligned} \hat{\mathbf{X}}_m &= \mathbf{X}_m \odot \mathbf{a}, \\ \mathbf{a} &= g(\mathbf{X}_m; \Lambda), \end{aligned} \tag{3}$$

where $\odot$ indicates the element-wise product[2], $\mathbf{a}$ is the learned importance weight vector, $\hat{\mathbf{X}}_m$ is the obtained weighted feature volume, and Λ is the network parameters in the MAFE module $g(\cdot)$.

By using the weighted features $\hat{\mathbf{X}}_m^{ET}, \hat{\mathbf{X}}_m^{TC}, \hat{\mathbf{X}}_m^{WT}$ to predict the segmentation masks of the interested tumor areas, we obtain:

$$\begin{cases} P(\hat{y}_{m,i}^{ET} = 1 | \hat{\mathbf{X}}_m^{ET}) = F_I(\mathbf{X}_m; \Phi_I,); \\ P(\hat{y}_{m,i}^{TC} = 1 | \hat{\mathbf{Y}}_m^{ET}, \hat{\mathbf{X}}_m^{TC}) = F_{II}(\hat{\mathbf{Y}}_m^{ET}, \mathbf{X}_m; \Phi_{II}); \\ P(\hat{y}_{m,i}^{WT} = 1 | \hat{\mathbf{Y}}_m^{TC}, \hat{\mathbf{X}}_m^{WT}) = F_{III}(\hat{\mathbf{Y}}_m^{TC}, \mathbf{X}_m; \Phi_{III}), \end{cases} \tag{4}$$

where $\Phi_I = \{\mathbf{W}_I, \Lambda_I\}$, $\Phi_{II} = \{\mathbf{W}_{II}, \Lambda_{II}\}$, and $\Phi_{III} = \{\mathbf{W}_{III}, \Lambda_{III}\}$ indicate the parameters of the three successive network branches $F_I(\cdot)$, $F_{II}(\cdot)$, and $F_{III}(\cdot)$, respectively. Here each of the network branches contains a MAFE module and an inferring module.

D. Objective Function and Training Strategy

The learning process of the proposed network model minimizes two-fold loss functions. The first one is the mean Dice coefficient loss. Compared with the cross-entropy loss or mean square error loss, the Dice coefficient loss can alleviate the imbalance issue of the training data in different classes [22], which fits to the brain tumor segmentation task well. Specifically, given each predicted segmentation mask with N voxels $\{\widehat{y}_i\}_{i=1}^N$ and the corresponding ground-truth segmentation mask $\{y_i\}_{i=1}^N$, the Dice coefficient loss is defined as follows:

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^N \delta(y_i = 1)\widehat{y}_i}{\sum_{i=1}^N [\delta(y_i = 1)]^2 + \sum_{i=1}^N \widehat{y}_i^2}, \tag{5}$$

where $\delta(\cdot)$ is the logical operator.

The other loss function is inspired by the two-sample t-test process [27]. As we know, t-test is often used to determine if two sets of data are significantly different from each other. Thus, in our approach, we use the t-test to measure if the predicted likelihood values in the tumor area are significantly different from the values in the non-tumor area. Denote the mean value of the predicted likelihood values in the tumor area as μ_f, the mean value of the predicted likelihood values in the non-tumor area as μ_b, the voxel numbers of the tumor area and non-tumor area as N_f and N_g, respectively. Then, if $t = (\mu_f - \mu_b)/\sqrt{\frac{S_f^2}{N_f} + \frac{S_g^2}{N_g}} \geq \epsilon$, where S_f^2 and S_g^2 are the variances of the predicted likelihood values in the tumor area and non-tumor area, respectively, we can hold the hypothesis that $\mu_f > \mu_g$. In our task, since we need to separate the tumor area from the non-tumor area, we can maximize t to encourage the network to predict high values in the tumor area but low values in the non-tumor area. To maximize t approximately, we optimize the following object function:

$$L_{t-test} = (\mu_f - \mu_b) - (\frac{S_f^2}{N_f} + \frac{S_g^2}{N_g}). \tag{6}$$

Then, the final loss function for segmenting each tumor area becomes:

$$L = L_{Dice} - \alpha L_{t-test}, \tag{7}$$

where α is a free parameter to weigh the t-test loss during the learning process.

During the training process, we find it a little bit hard to train the whole network from scratch. Instead, we implement a two-stage learning procedure. Specifically, we split the whole network into three parts, each of which contains a MAFE module and an inferring module to segment a certain type of tumor area. Then, we pre-train each of the sub-networks according to the corresponding ground-truth annotation. After this pre-training stage, we train the parameters of the whole network by integrating these sub-networks and fine-tuning their parameters.

IV. Experiments

A. Data and Implementation Details

We use the BraTS 2017 and 2018 [32], [33], [34] benchmarks for experiments. The BraTS 2017 training set contains 3D volume data from 285 patients, among which 210 are high-grade gliomas (HGG) data while 75 are low-grade gliomas (LGG) data. The BraTS 2017 validation set contains 3D volume data from 46 patients with brain tumors of unknown grade. The BraTS 2018 training set also contains 3D volume data from 285 patients, among which 210 are HGG data while 75 are LGG data. The BraTS 2018 validation set contains 3D volume data from 66 patients with brain tumors of unknown grade[3]. The 3D volume data of each patient contains four modalities, which are the T1, T1c, T2 and FLAIR, respectively. These data have been skull-striped, re-sampled, and co-registered well. The ground truth data are segmentation masks manually

[2]During calculation, we will first extend $\mathbf{a}$ from vector to volume tensor.

[3]Notice that although the BraTS 2017 and 2018 have the same training set, we report the experimental results on their validation sets separately in order to be consistent with the evaluation in previous works.

TABLE I

Comparison of the proposed approach and other state-of-the-art methods on the BraTS 2017 in terms of the Dice score (the higher the better), Hausdorff distance (the lower the better), and model parameter (the lower the better). Besides the absolute number of model parameters of each compared method, we also report the ratio (under the number of parameters of each model) when comparing the model parameter of our approach to that of the other method.

		Enhancing Tumor		Whole Tumor		Tumor Core		Average		Model
		Dice	Hausdorff	Dice	Hausdorff	Dice	Hausdorff	Dice	Hausdorff	Parameter
Lopez et al. [20]	Mean:	0.567	23.828	0.783	30.316	0.685	38.077	0.678	30.740	1.57E7
	StdDev:	-	-	-	-	-	-	-	-	2.31%
Shaikh et al. [18]	Mean:	0.650	-	0.870	-	0.680	-	0.733	-	2.31E7
	StdDev:	0.320	-	0.110	-	0.340	-	0.257	-	1.57%
Alex et al. [15]	Mean:	0.690	-	0.830	-	0.690	-	0.737	-	8.63E6
	StdDev:	0.320	-	0.160	-	0.300	-	0.260	-	4.20%
Castillo et al. [23]	Mean:	0.690	-	0.860	-	0.690	-	0.747	-	2.22E7
	StdDev:	-	-	-	-	-	-	-	-	1.63%
Li et al. [22]	Mean:	0.704	7.699	0.871	10.396	0.682	13.062	0.752	10.386	3.80E5
	StdDev:	0.307	14.407	0.083	15.754	0.304	17.573	0.231	15.911	95.4%
Islam et al. [19]	Mean:	0.689	12.938	0.876	9.82	0.761	12.361	0.775	11.706	1.34E8
	StdDev:	0.304	26.453	0.086	13.516	0.221	20.826	0.204	20.265	0.27%
Andermatt et al. [28]	Mean:	0.711	4.187	0.893	4.613	0.735	8.189	0.780	**5.663**	-
	StdDev:	0.304	6.112	0.086	5.732	0.299	13.813	0.230	8.552	-
Jesson et al. [29]	Mean:	0.713	6.980	0.899	4.160	0.751	8.650	0.788	6.597	4.29E6
	StdDev:	0.291	12.100	0.070	3.370	0.240	9.350	0.200	**8.273**	8.45%
Havaei et al. [30]	Mean:	0.730	-	0.880	-	0.790	-	0.800	-	8.02E5
	StdDev:	-	-	-	-	-	-	-	-	45.2%
Wang et al. [31] (S)	Mean:	0.740	5.318	0.890	12.457	0.820	9.662	0.817	9.146	5.95E5
	StdDev:	-	-	-	-	-	-	-	-	60.9%
OURS	Mean:	0.766	4.147	0.883	8.081	0.818	10.059	**0.822**	7.429	**3.63E5**
	StdDev:	0.267	5.557	0.084	11.154	0.154	9.583	**0.169**	8.765	-

annotated by experts[4]. To evaluate the segmentation accuracy, we adopt the widely used Dice score and Hausdorff distance. As the validation sets do not provide the ground-truth annotation masks, we obtain the evaluation scores from their evaluation website.

We implement our network in Tensorflow [35]. The training process is implemented on a NVIDIA GTX 1080Ti GPU. We adopt the Adaptive Moment Estimation (Adam) [36] for training, with an initial learning rate 10^{-4}, weight decay 10^{-7}, batch size 2, and maximal iteration 140k. α is set to 0.1. We take in total 29 hours to train and 10.6s per volume to test. Our network has 3.5e5 learnable parameters which are less than some state-of-the-art brain tumor segmentation networks like Havaei [30] (8.0e5), indicating that our model has moderate complexity. For pre-processing, we follow [31] to adopt a very simple operation, which normalizes each image of the 3D volume data by its mean value and standard deviation. For post-processing, we remove small isolated areas to correct some voxel labels using a simple thresholding method. The threshold is set as half of the number of pixels residing in the biggest connected area in each predicted binary map. In addition, we also remove noisy areas that are smaller than 500 pixels when predicting the ET areas.

B. Comparison to the State-of-the-arts

In this subsection, we compare the proposed approach with 13 state-of-the-art methods, i.e., [20], [18], [15], [23], [22], [19], [28], [29], [30], [37], [38], [39], [40], which mainly use the U-net-like network architectures like ours and have the comparable scales of network parameters to our approach. These methods also include both the 2D CNN-based methods and 3D CNN-based methods. Table I and Table II report the comparison results on the BraTS 2017 and 2018 validation set, respectively, from which we can observe that the proposed approach achieves the superior performance when compared with the existing state-of-the-art methods. Specifically, in terms of the mean Dice score, our approach achieves 0.822 and 0.834 on the two datasets, which outperforms the other state-of-the-art methods by 2.2% to 14.4% and 0.7% to 10.1%, respectively. Besides, in terms of the standard deviation of the Dice score, our approach achieves 0.169 and 0.156 on the two datasets, which is also better than the other state-of-the-art methods. It is worth mentioning that the Wang et al. [31] (S) indicates the single-view model of [31] (more specifically the axial view), which has a more comparable setting with our approach as our model is also trained on the data along the axial view[5]. Based on the comparison results between Wang et al. [31] (S) and our approach, we can observe that by exploring the task structure, our approach is able to obtain slightly better segmentation accuracy while reducing nearly 40% of the network parameters.

Seen from the comparison results with [24] and [25], it is clear that our approach obtains superior performance for all the three tumor areas under both the Dice score and the Hausdorff distance. To our best knowledge, there are two

[4]The detailed information about the dataset and ground-truth can be referred to in https://www.med.upenn.edu/sbia/brats2018/data.html

[5]Training the model along other views would obtain different performance. When training along the coronal view and sagittal view, our model obtains 0.814 and 0.823 mean Dice scores on the BraTS 2018 validation set, respectively.

TABLE II

Comparison of the proposed approach and other state-of-the-art methods on the BraTS 2018 in terms of the Dice score (the higher the better), Hausdorff distance (the lower the better), and model parameter (the lower the better). Besides the absolute number of the model parameter of each compared method, we also report the ratio (under the number of parameters of each model) when comparing the model parameters of our approach to those of the other methods. Notice that as [25] and [24] do not have the open access projects and results, we implement their algorithms ourselves based on their descriptions for experimental comparison.

		Enhancing Tumor		Whole Tumor		Tumor Core		Average		Model
		Dice	Hausdorff	Dice	Hausdorff	Dice	Hausdorff	Dice	Hausdorff	Parameter
Carver et al. [37]	Mean:	0.710	4.460	0.880	7.090	0.770	9.570	0.787	7.040	2.20E7
	StdDev:	0.290	8.320	0.080	11.570	0.260	14.080	0.210	11.323	1.65%
Benson et al. [38]	Mean:	0.660	15.940	0.820	26.410	0.720	18.870	0.733	20.407	-
	StdDev:	0.270	25.560	0.100	23.610	0.230	20.560	0.200	23.243	-
Chen et al. [25]	Mean:	0.707	10.385	0.845	11.822	0.731	15.066	0.761	12.424	1.33E7
	StdDev:	0.264	21.205	0.116	19.037	0.245	19.810	0.208	20.017	2.73%
Salehi et al. [24]	Mean:	0.704	9.668	0.822	9.610	0.733	13.909	0.753	11.062	1.20E7
	StdDev:	0.289	13.757	0.136	13.036	0.242	14.965	0.222	13.919	3.03%
Puch et al. [39]	Mean:	0.758	4.502	0.895	10.656	0.774	7.103	0.809	8.880	1.48E6
	StdDev:	0.264	8.227	0.070	19.286	0.253	7.084	0.196	11.532	24.53%
Chandra et al. [40]	Mean:	0.767	7.569	0.901	6.680	0.813	7.630	0.827	7.293	-
	StdDev:	-	-	-	-	-	-	-	-	-
Isensee et al. [41]	Mean:	0.807	2.74	0.909	5.83	0.852	7.20	0.856	**5.2567**	1.45E7
	StdDev:	-	-	-	-	-	-	-	-	2.50%
Myronenko et al. [42]	Mean:	0.815	3.8048	0.904	4.4834	0.860	8.2777	**0.859**	5.5220	2.01E7
	StdDev:	-	-	-	-	-	-	-	-	1.81%
OURS	Mean:	0.782	3.567	0.896	5.733	0.824	9.270	0.834	6.190	**3.63E5**
	StdDev:	0.232	4.286	0.062	7.665	0.173	13.238	**0.156**	**8.396**	-

reasons. First, our approach pays special attention to the mask-modality structure and uses the t-test loss for brain tumor segmentation. Second, both [24] and [25] process 2D convolution on each slice separately, while our approach performs 3D convolution on voxels which explores the richer context for brain tumor segmentation. In addition, compared with the state-of-the-art methods with heavy networks, such as [41] and [42], our method can obtain approaching performance (less than 2.5% performance gap in terms of the Mean Dice score) by only having about 0.25% parameters of them (i.e., x40 reduction in memory cost). This also demonstrates the effectiveness of our approach to some extent and implies that the proposed network would have better potential in applications on devices with limited computing capacity and memory. In summary, the experimental results in Table I and Table II demonstrate that our proposed approach is able to achieve precise and robust brain tumor segmentation results. The promising performance of our approach is obtained mainly from the strategy to explore the task-structure in network design as the used network layers or blocks are well-established ones. In this way, the value of our learning strategy is better demonstrated.

C. Ablation Study of the Proposed Approach

In this subsection, we carry out ablation studies of the proposed approach to evaluate and analyze the components we have considered. Specifically, we compare our approach with the following seven baselines:

- **OURS w/o TS&MAFE:** Removing the task-structured prediction mechanism and the modality-aware embedding mechanism from our proposed network, which is the most basic baseline for our approach.
- **OURS w/o TS:** Removing the task-structured prediction mechanism from our proposed network, i.e., using the same network architecture but simultaneously predicting the three tumor areas at the end of the network. Notice that this baseline contains a MAFE module followed by three cascaded inferring modules, where the last conv layers in the first two inferring modules are removed and the last conv layer in the last inferring module predicts three tumor areas, simultaneously. As

TABLE III

Comparison of the proposed approach and other baseline models in terms of the Dice score, where "ET", "WT", "TC", and "StdDev" are short for "Enhancing Tumor", "Whole Tumor", "Tumor Core", and "Standard deviation", respectively.

		ET	WT	TC	Average
OURS w/o TS&MAFE	Mean:	0.704	0.871	0.682	0.752
	StdDev:	0.307	0.083	0.304	0.231
OURS w/o TS	Mean:	0.718	0.877	0.764	0.786
	StdDev:	0.298	0.071	0.226	0.198
OURS w/o MAFE	Mean:	0.748	0.892	0.740	0.794
	StdDev:	0.293	0.066	0.241	0.200
OURS w/o FE	Mean:	0.740	0.891	0.729	0.787
	StdDev:	0.289	0.089	0.291	0.223
OURS w/o PT	Mean:	0.764	0.896	0.751	0.804
	StdDev:	0.267	0.063	0.232	0.187
OURS w/o t-test	Mean:	0.763	0.891	0.731	0.795
	StdDev:	0.269	0.071	0.275	0.205
OURS L2S	Mean:	0.731	0.866	0.725	0.774
	StdDev:	0.291	0.089	0.244	0.208
OURS	Mean:	0.766	0.883	0.818	**0.822**
	StdDev:	0.267	0.084	0.154	**0.169**

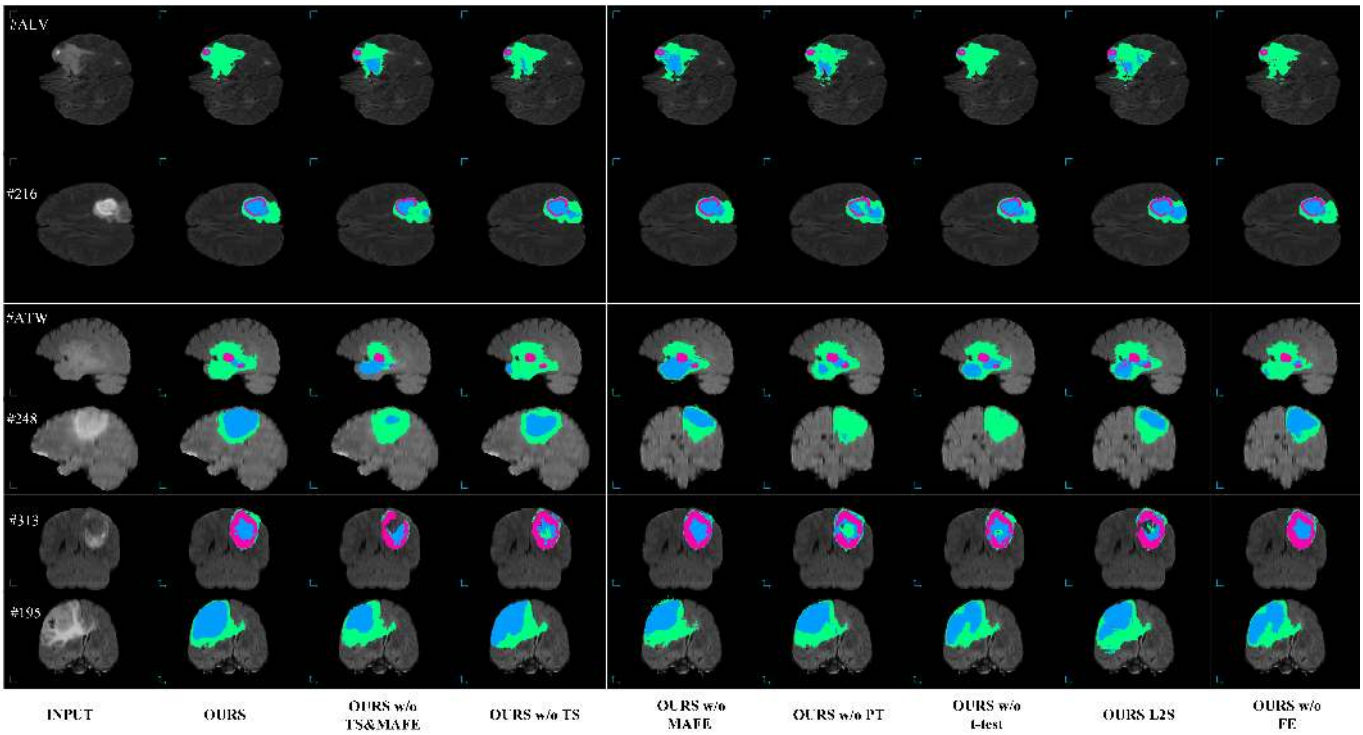

Fig. 6: Examples of the brain tumor segmentation results from the BraTS 2017 Validation Dataset. The green, blue, and pink regions indicate the whole tumor areas, tumor core areas, and enhancing tumor core areas, respectively.

a result, this baseline has almost the same network depth as ours.

- **OURS w/o MAFE:** Removing the MAFE modules from our proposed network architecture.
- **OURS w/o FE:** Directly using the feature volume that is obtained by concatenating the different modality features with equal importance weights (see the left part of Fig. 4) as the output of the MAFE module.
- **OURS w/o PT:** Learning the network in one stage without the pre-training stage.
- **OURS w/o t-test:** Only using the Dice loss during the training process of the proposed network.
- **OURS L2S:** Using the proposed network architecture to segment tumor areas from the largest one (i.e., the WT area) to the smallest one (i.e., the ET area).

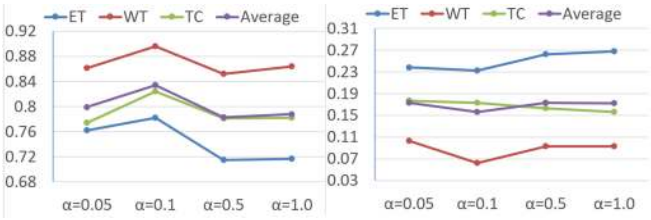

Fig. 7: Influence of the parameter α on the BraTS 2018 dataset. The left figure is based on the measurement of the mean Dice score, where higher values indicate better results. The right figure is based on the measurement of the standard deviation of the Dice score, where lower values indicate better results.

The experimental results are reported in Table III. The comparison between **OURS w/o TS&MAFE** and OURS demonstrates that the components considered in our approach can significantly improve the performance (7.0% in terms of the mean Dice score) over the commonly used baseline network. The comparison of **OURS w/o TS** to **OURS w/o MAFE** and OURS demonstrates that both the task-structured prediction and MAFE are important for this task, while the former plays a more important role in improving the performance. The comparison of **OURS w/o MAFE** to **OURS w/o FE** implies that simply using more convolutional layers without the task-modality weighting mechanism won't always help learn better features. While the comparison between **OURS w/o FE** and OURS demonstrates the effectiveness of using the weighted combination strategy to model the task-modality structure. The comparison of **OURS w/o PT** to **OURS w/o t-test** and OURS demonstrates that the proposed t-test loss and the two-stage learning strategy also bring benefit to our approach. The comparison of **OURS L2S** to OURS demonstrates that the adopted small-to-large task structure obtains better performance than the large-to-small task structure, which is consistent with the prior knowledge in clinical practice (see in Sec. I). The results obtained by **OURS w/o TS**, **OURS w/o MAFE**, **w/o t-test**, and OURS are visualized in Fig. 6.

From the above experiments, we observe an interesting phenomenon that although the overall performance of our approach is always better than the other baseline methods, the performance on the WT area is slightly worse than some baseline methods. Based on our investigation, this

is caused by the loss-imbalance issue. Compared to the ET and TC areas, the WT area is larger, leading to its Dice loss value much smaller than its t-test loss value. This makes the optimization regarding the WT area mainly depend on the t-test loss, thus limiting the test performance which is under the measurement of Dice. This is why using t-test loss does not help segment the WT area. This issue also prevents us from learning optimal parameters in the MAFE or the FE module.

Besides, to study the influence of using different α values to weigh the t-test loss, we further implement the sensitivity study on α. The experimental results are shown in Fig. 7, from which we can observe that the performance of our approach is sensitive to α and would reach a peak value when $\alpha = 0.1$.

To show the task-modality structure that is discovered by our network, we perform the statistical analysis on the importance weights that are inferred from the MAFE module. Specifically, we calculate the average values of the inferred importance weights for each modality data when segmenting a certain type of tumor area and report them in Fig. 8. As can be seen, our network learns that the T1-contrast, T1, and FLAIR modality contain the most informative cues when segmenting the ET, TC, and WT areas, respectively. From Fig. 1, we can observe that when we conduct this task, the T1-contrast modality is indeed more useful to segment the ET area. This demonstrates the effectiveness of the task-modality structure learned by our TSBTS net.

Finally, we also visualize several failure cases of the proposed approach. As shown in Fig. 9, the failure cases may appear when the input image has a small contrast between the tumor area and the background area (see the first column), the tumor areas have complex shapes and appearances (see the middle columns) or even do not present any observable structure (see the last column). Besides, from the examples in the last two columns of Fig. 9, we can observe that our approach can obtain good performance on the WT when the segmentation of the TC and/or ET area is a failure. This indicates that in the proposed learning framework, the errors in the former inferring module won't dramatically affect the learning process of the latter inferring module.

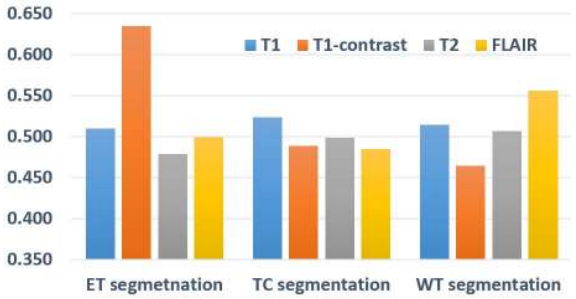

Fig. 8: The average values of the importance weights for each modality data inferred by our TSBTS net when segmenting a certain type of tumor area.

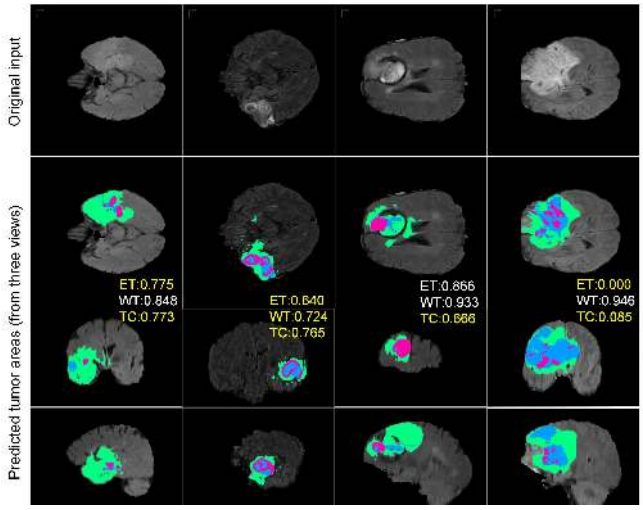

Fig. 9: Some examples of the failure cases of our approach. The first row shows the original input with the FLAIR modality. The green, blue, and pink regions indicate the WT areas, TC areas, and ET areas, respectively. The scores reported in the middle of each column are the Dice scores of the predicted tumor areas. Here scores marked in yellow (i.e., those below 0.8) are considered as the failure cases.

V. Conclusion

In this paper, we have proposed a novel deep neural network model to explore task structure and modality importance for multi-modality brain tumor segmentation. This is based on two findings: On one hand, the three targeted tumor areas are mutually included rather than being located separately. On the other hand, different modalities are of different importance for segmenting tumor areas. For implementing the task-structured learning, we predict the different types of brain tumor areas in different network modules. For exploring the modality importance, we introduce the modality-aware feature embedding mechanism to our network to infer the importance weights and the weighted features. Comprehensive experiments have demonstrated the effectiveness of the components considered in our approach as well as the outstanding capacity of our entire approach when compared with the state-of-the-art methods. In future works, we will explore the task structure for other tasks in the medical image analysis field or the conventional image or video understanding fields and apply our approach to solve those problems. We will also explore the potential of integrating the saliency detection technique [43], [44], [45] and the weakly supervised learning technique [46], [47] to further improve learning performance.

VI. Acknowledgement

This work was supported in part by the National Science Foundation of China under Grants 61876140 and 61773301, and the China Postdoctoral Support Scheme for Innovative Talents under Grant BX20180236.

REFERENCES

[1] M. Kampffmeyer, N. Dong, X. Liang, Y. Zhang, and E. P. Xing, "Connnet: A long-range relation-aware pixel-connectivity network for salient segmentation," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2518–2529, May 2019.

[2] J. Fu, J. Liu, Y. Wang, J. Zhou, C. Wang, and H. Lu, "Stacked deconvolutional network for semantic segmentation," *IEEE Transactions on Image Processing*, pp. 1–1, 2019.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[4] Y. Fang, G. Ding, J. Li, and Z. Fang, "Deep3dsaliency: Deep stereoscopic video saliency detection model by 3d convolutional networks," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2305–2318, May 2019.

[5] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1992–2004, April 2017.

[6] D. Zikic, B. Glocker, E. Konukoglu, A. Criminisi, C. Demiralp, J. Shotton, O. M. Thomas, T. Das, R. Jena, and S. J. Price, "Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel mr," in *MICCAI*. Springer, 2012, pp. 369–376.

[7] R. Meier, S. Bauer, J. Slotboom, R. Wiest, and M. Reyes, "Appearance-and context-sensitive features for brain tumor segmentation," *Proceedings of MICCAI BRATS Challenge*, pp. 020–026, 2014.

[8] R. Meier, S. Bauer, J. Slotboom, and M. Reyes, "A hybrid model for multimodal brain tumor segmentation," *Multimodal Brain Tumor Segmentation*, vol. 31, pp. 31–37, 2013.

[9] N. J. Tustison, K. Shrinidhi, M. Wintermark, C. R. Durst, B. M. Kandel, J. C. Gee, M. C. Grossman, and B. B. Avants, "Optimal symmetric multimodal templates and concatenated random forests for supervised brain tumor segmentation (simplified) with antsr," *Neuroinformatics*, vol. 13, no. 2, pp. 209–225, 2015.

[10] C.-H. Lee, S. Wang, A. Murtha, M. R. Brown, and R. Greiner, "Segmenting brain tumors using pseudo–conditional random fields," in *MICCAI*. Springer, 2008, pp. 359–366.

[11] S. Bauer, L.-P. Nolte, and M. Reyes, "Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization," in *MICCAI*. Springer, 2011, pp. 354–361.

[12] A. Pinto, S. Pereira, H. Correia, J. Oliveira, D. M. Rasteiro, and C. A. Silva, "Brain tumour segmentation based on extremely randomized forest with high-level features," in *EMBC*. IEEE, 2015, pp. 3037–3040.

[13] E. Geremia, B. Menze, and N. Ayache, "Spatially adaptive random forest," in *ISBI*. IEEE, 2013, pp. 1332–35.

[14] R. Meier, S. Bauer, J. Slotboom, R. Wiest, and M. Reyes, "Patient-specific semi-supervised learning for postoperative brain tumor segmentation," in *MICCAI*. Springer, 2014, pp. 714–721.

[15] M. S. Varghese Alex and G. Krishnamurthi, "Brain tumor segmentation from multi modal mr images using fully convolutional neural network," in *International MICCAI Brainlesion Workshop*. Springer, 2017, pp. 1–8.

[16] S. Pereira, V. Alves, and C. A. Silva, "Adaptive feature recombination and recalibration for semantic segmentation: application to brain tumor segmentation in mri," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 706–714.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[18] M. Shaikh, G. Anand, G. Acharya, A. Amrutkar, V. Alex, and G. Krishnamurthi, "Brain tumor segmentation using dense fully convolutional neural network," in *International MICCAI Brainlesion Workshop*. Springer, 2017, pp. 309–319.

[19] M. Islam and H. Ren, "Fully convolutional network with hypercolumn features for brain tumor segmentation," in *Proceedings of MICCAI workshop on Multimodal Brain Tumor Segmentation Challenge (BRATS)*, 2017.

[20] M. M. Lopez and J. Ventura, "Dilated convolutions for brain tumor segmentation in mri scans," in *International MICCAI Brainlesion Workshop*. Springer, 2017, pp. 253–262.

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

[22] W. Li, G. Wang, L. Fidon, S. Ourselin, M. J. Cardoso, and T. Vercauteren, "On the compactness, efficiency, and representation of 3d convolutional networks: brain parcellation as a pretext task," in *IPMI*. Springer, 2017, pp. 348–360.

[23] L. S. Castillo, L. A. Daza, L. C. Rivera, and P. Arbeláez, "Volumetric multimodality neural network for brain tumor segmentation," in *13th International Conference on Medical Information Processing and Analysis*, vol. 10572. International Society for Optics and Photonics, 2017, p. 105720E.

[24] S. S. Mohseni Salehi, D. Erdogmus, and A. Gholipour, "Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging," *IEEE Transactions on Medical Imaging*, vol. 36, no. 11, pp. 2319–2330, Nov 2017.

[25] X. Chen, J. H. Liew, W. Xiong, C.-K. Chui, and S.-H. Ong, "Focus, segment and erase: An efficient network for multi-label brain tumor segmentation," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 674–689.

[26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.

[27] B. Efron and T. Hastie, *Computer age statistical inference*. Cambridge University Press, 2016, vol. 5.

[28] S. P. l. Simon Andermatt and P. Cattin, "Multi-dimensional gated recurrent units for brain tumor segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2017, pp. 15–19.

[29] A. Jesson and T. Arbel, "Brain tumor segmentation using a 3d fcn with multi-scale loss," in *International MICCAI Brainlesion Workshop*. Springer, 2017, pp. 392–402.

[30] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P. M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical Image Analysis*, vol. 35, pp. 18–31, 2017.

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

[32] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, p. 1993, 2015.

[33] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos, "Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection," *The Cancer Imaging Archive*, vol. 286, 2017.

[34] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, p. 170117, 2017.

[35] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.

[36] D. P. Kingma and J. L. Ba, "Adam: Amethod for stochastic optimization," in *ICLR*, 2014.

[37] E. Carver, C. Liu, W. Zong, Z. Dai, J. M. Snyder, J. Lee, and N. Wen, "Automatic brain tumor segmentation and overall survival prediction using machine learning algorithms," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 406–418.

[38] E. Benson, M. P. Pound, A. P. French, A. S. Jackson, and T. P. Pridmore, "Deep hourglass for brain tumor segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 419–428.

[39] S. Puch, I. Sánchez, A. Hernández, G. Piella, and V. Prc?kovska, "Global planar convolutions for improved context aggregation in brain tumor segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 393–405.

[40] S. Chandra, M. Vakalopoulou, L. Fidon, E. Battistella, T. Estienne, R. Sun, C. Robert, E. Deutsch, and N. Paragios, "Context aware 3d cnns for brain tumor segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 299–310.

[41] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "No new-net," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 234–244.

[42] A. Myronenko, "3d mri brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 311–320.
[43] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: a survey," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 84–100, 2018.
[44] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *CVPR*, 2018, pp. 714–722.
[45] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *ECCV*, 2018, pp. 186–202.
[46] D. Zhang, J. Han, L. Zhao, and D. Meng, "Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework," *International Journal of Computer Vision*, vol. 127, no. 4, pp. 363–380, 2019.
[47] Y. Shen, R. Ji, C. Wang, X. Li, and X. Li, "Weakly supervised object detection via object-specific pixel gradient," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 12, pp. 5960–5970, 2018.

Dingwen Zhang received his Ph.D. degree from the Northwestern Polytechnical University, Xi'an, China, in 2018. He is currently an associate professor in School of Mechano-Electronic Engineering, Xidian University. From 2015 to 2017, he was a visiting scholar at the Robotic Institute, Carnegie Mellon University. His research interests include computer vision and multimedia processing, especially on saliency detection, video object segmentation, and weakly supervised learning.

Guohai Huang received his Bachelor degree from Xidian University, Xian, China, in 2017. He is currently pursuing the M.S. degree in the School of Mechano-Electronic Engineering, Xidian University. His research interests include pattern recognition and medical image processing.

Qiang Zhang received the B.S. degree in automatic control, the M.S. degree in pattern recognition and intelligent systems, and the Ph.D. degree in circuit and system from Xidian University, China, in 2001,2004, and 2008, respectively. He was a Visiting Scholar with the Center for Intelligent Machines, McGill University, Canada. He is currently a professor with the Automatic Control Department, Xidian University, China. His current research interests include image processing, pattern recognition.

Jungong Han is currently a Full Professor and Chair in Computer Science at Aberystwyth University, UK. His research interests span the fields of video analysis, computer vision and applied machine learning. He has published over 180 papers, including 50+ IEEE/ACM Transactions and 40+ A* conference papers.

Junwei Han (M'12-SM'15) is a Professor with Northwestern Polytechnical University, Xi'an, China. He received Ph.D. degree in Northwestern Polytechnical University in 2003. He was a Research Fellow in Nanyang Technological University, The Chinese University of Hong Kong, and University of Dundee. His research interests include computer vision and brain imaging analysis. He has published over 100 papers in IEEE TRANSACTIONS and top tier conferences. He is currently an Associate Editor of IEEE Trans. on Human-Machine Systems, Neurocomputing, and Machine Vision and Applications.

Yizhou Wang Yizhou Wang received the bachelors degree in electrical engineering from Tsinghua University, in 1996 and the PhD degree in computer science from the UCLA, in 2005. He is a professor of Computer Science Department, Peking University, Beijing, China. His research interests include computational vision, statistical modeling and learning, pattern analysis, and digital visual arts.

Yizhou Yu (M'10-SM'12-F'19) received the Ph.D. degree from the University of California at Berkeley in 2000. He is currently a Professor with The University of Hong Kong. He was a Faculty Member at the University of Illinois at UrbanaCChampaign for 12 years. His current research interests include computer vision, deep learning, biomedical data analysis, computational visual media, and geometric computing. He is a recipient of the 2002 U.S. National Science Foundation CAREER Award, the 2007 NNSF China Overseas Distinguished Young Investigator Award, and the ACCV 2018 Best Application Paper Award. He has served on the editorial boards of the IET Computer Vision, The Visual Computer, and the IEEE Transactions on Visualization and Computer Graphics. He has also served on the program committees of many leading international conferences, including SIGGRAPH, SIGGRAPH Asia, and the International Conference on Computer Vision.