CANet: Context Aware Network for Brain Glioma Segmentation

Zhihua Liu, Lei Tong, Long Chen, Feixiang Zhou, Zheheng Jiang, Qianni Zhang, Yinhai Wang, Caifeng Shan, *Senior Member, IEEE,* Ling Li and Huiyu Zhou

***Abstract*—Automated segmentation of brain glioma plays an active role in diagnosis decision, progression monitoring and surgery planning. Based on deep neural networks, previous studies have shown promising technologies for brain glioma segmentation. However, these approaches lack powerful strategies to incorporate contextual information of tumor cells and their surrounding, which has been proven as a fundamental cue to deal with local ambiguity. In this work, we propose a novel approach named Context-Aware Network (CANet) for brain glioma segmentation. CANet captures high dimensional and discriminative features with contexts from both the convolutional space and feature interaction graphs. We further propose context guided attentive conditional random fields which can selectively aggregate features. We evaluate our method using publicly accessible brain glioma segmentation datasets BRATS2017, BRATS2018 and BRATS2019. The experimental results show that the proposed algorithm has better or competitive performance against several State-of-The-Art approaches under different segmentation metrics on the training and validation sets.**

***Index Terms*—Brain glioma, conditional random field, graph convolutional network, image segmentation.**

I. Introduction

GLIOMA is one of the most prevalent types of adult brain tumor with fateful health damage impacts and high mortality [1]. To provide sufficient evidence for early diagnosis, surgery planning and post-surgery observation, Magnetic Resonance Imaging (MRI) with multi-modalities (e.g. T1, T1 with contrast-enhanced (T1ce), T2 and Fluid Attenuation Inversion Recover (FLAIR)) is a widely used diagnosis technique to provide reproducible and non-invasive measurement, including structural, anatomical and functional characteristics.

Medical image segmentation provides fundamental guidance and quantitative assessment for medical professionals to achieve disease diagnosis, tumor growth monitoring, planning treatment and follow-up services [2], [3]. Fig. 1 shows an overview of the brain glioma segmentation task. However, manual segmentation requires professional expertise and tends to be time consuming and labour intensive. Previous methods on automated brain glioma segmentation were based on traditional machine learning algorithms [4]–[7], which strongly rely on hand-crafted features, such as textures [8] and local histograms [9]. However, finding the best hand-crafted features or optimal feature combinations is impracticable. In recent years, deep learning techniques, especially deep convolutional neural network (DCNN), have been deployed to effectively learn high dimensional discriminative features from data and widely used on various medical imaging tasks [10].

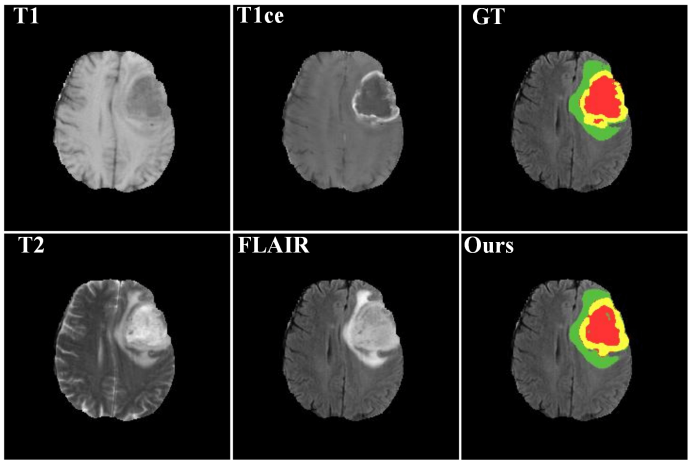


Fig. 1. Examples of multi-modality image slices from BraTS17 with the ground-truth and our segmentation results. In this figure, green represents GD-Enhancing Tumor (numerical label 2), yellow represents Pertumoral Edema (numerical label 1) and red represents Necrotic and Non-Enhancing Tumor Core (NCR\ECT, numerical label 4).

Inter-class ambiguity is a common issue in brain glioma segmentation. This issue makes it hard to achieve accurate dense voxel-wise segmentation if we only consider isolated voxels, as different classes' voxels may share similar intensity values or feature representations. To address this issue, we aim to learn relational information between glioma cells and their surroundings by exploring their feature interaction graphs. We here propose a context-aware network, namely CANet, to achieve accurate dense voxel-wise brain glioma segmentation in MRI images. Our contributions in this work are summarised below:

- We propose a novel brain glioma segmentation approach by introducing feature interaction graph reasoning as a parallel auxiliary branch to model the relation between

Zhihua Liu, Lei Tong, Long Chen, Feixiang Zhou, Zheheng Jiang are with the School of Informatics, University of Leicester, Leicester, United Kingdom, LE1 7RH. E-mail: zl208@leicester.ac.uk

Qianni Zhang is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom

Yinhai Wang is with Biopharmaceutical R&D, AstraZeneca, Unit 310 - Darwin Building, Cambridge Science Park, Milton Road, Cambridge, United Kingdom, CB4 0WG

Caifeng Shan is with Philips Research, High Tech Campus, 5656 AE, Eindhoven, The Netherlands

Ling Li is with the School of Computing, University of Kent, United Kingdom

Huiyu Zhou is with the School of Informatics, University of Leicester, Leicester, United Kingdom, LE1 7RH. Corresponding author, E-mail: hz143@leicester.ac.uk

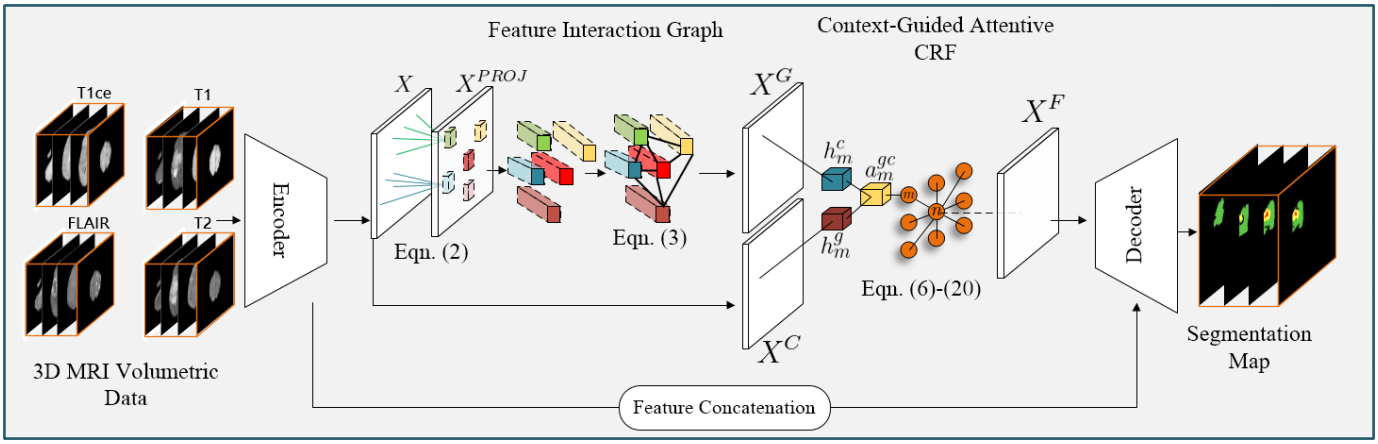


Fig. 2. The architecture of the proposed context aware network.

glioma cells and their surroundings. The intermediate feature representations are further exploited and aggregated within a customized context guided attentive conditional random field (CGA-CRF) framework. To our knowledge, this is the first practice on brain glioma segmentation that incorporates relational information from the generated features.

- We formulate the mean-field approximation of the inferences in the proposed CGA-CRF as a convolution operation, whereas CGA-CRF is implemented as sequential deep neural networks layers. Our formulation demonstrates the generalization capability of the proposed CGA-CRF that can be embedded within any deep neural architecture seamlessly to achieve end-to-end training.
- We conduct extensive evaluations to demonstrate that our proposed approach outperforms several State-of-The-Art methods under different evaluation metrics on the Multimodal Brain Tumor Image Segmentation Challenge (BraTS) datasets, i.e. BraTS2017, BraTS2018 and BraTS2019.

II. Related Work

We construct our novel brain glioma segmentation approach upon recent successes of deep neural networks and probabilistic graphical models. Below, we briefly review related methods categorising them into three sub-areas, *i.e.* brain glioma segmentation, semantic segmentation using conditional random field combined with convolutional neural network, and graph neural network in medical image analysis.

Brain Glioma Segmentation. Early research works on brain glioma segmentation mainly used traditional machine learning algorithms, such as clustering [6], random decision forests [7], Bayesian models [11] and graph-cuts [12]. Shin [6] used sparse coding for generating edema features and K-means for clustering tumor voxels. However, how to optimise the size of coding dictionary is an intractable problem. Pereira et al. [13] classified voxels' labels using random decision forests, which heavily relied on hand-crafted features and complicated postprocessing. Corso et al. [11] used a Bayesian formulation for incorporating soft model assignments into the affinity calculation. This method considered the weighted aggregation of multi-scale features, but ignored the relationship between different scales. Wels et al. [12] proposed a graph-cut based method to learn optimal graph representations for tumor segmentation, resulting in a superior performance. However, this method required a prolonged inference duration for dense segmentation tasks, as the number of vertices in its graph is proportional to the number of the voxels. Konukoglu et al. [14] and Menze et al. [15] incorporated a reaction-diffusion based biophysical tumor growth framework for glioma segmentation. The former focused on constructing the irradiation invasion margin at a single time instance while the latter focused on formalizing the macroscopic tumor growth model on longitude data. However, both methods required detailed domain knowledge to define the parameters, which limits generalization performance of their methods.

Promising achievements have been made on multi-modal MRI brain glioma segmentation using DCNN. Zikic et al. [16] was one of the earliest works that apply DCNN onto brain glioma segmentation. Havaei et al [17] proposed an improved DCNN by using muconvolutional kernels to extract local and global features. Zhao et al. [18] proposed a modified fully convolutional network (FCN) [19] with conditional random fields as post-processing module for refining brain glioma segmentation. Dong et al. [20] proposed a modified U-Net [21] for brain glioma segmentation. These previous works used 2D convolutional kernels on 2D image slices generated from the original 3D volumetric data. Methods of using 2D slices decrease the number of the used parameters and require less memory. However, this procedure also leads to the missing of spatial contexts. To minimise information loss and exploit the evidence of adjacent slices, Lyksborg et al. [22] ensembled three 2D CNNs on three orthogonal 2D patches.

To fully make use of 3D information, recent works applied 3D convolutional kernels onto volumetric data. Kamnitsas et al. [23] proposed a two pathway 3D DCNN, followed by a dense CRF, for brain glioma segmentation. Authors of [23] further extended the work using model ensembling [24]. Their proposed system EMMA ensembled the models from fully convolutional network (FCN) [19], U-Net [21] and

DeepMedic [25]. To avoid over-fitting problems in 3D voxel-level segmentation on limited training datasets, Myronenko [26] proposed a 3D DCNN with an additional variational autoencoder to regularise the decoder by reconstructing the input image.

Medical image datasets (e.g. BraTS) usually have imbalance and inter-class interference problems. To address this issue whilst maintaining segmentation performance, Chen et al. [27] and Wang et al. [28] both applied cascaded network structures for segmenting brain glioma, where the input of the inner region segmentation network is the output of the outer region segmentation network. However, cascaded cropping networks mainly focus on one tumor region in one particular network stage and cannot infer the relationship between different tumor regions.

Semantic segmentation using conditional random field and convolutional neural network. Brain glioma segmentation, along with the generic semantic segmentation, aims to assign each pixel with a specific semantic label. Among all the traditional machine learning methods, probabilistic graphical model, especially conditional random field has been considered as one of the most successful representation methods for solving semantic segmentation tasks [29]–[32]. In order to learn complex unary and pairwise potentials in the end-to-end fashion, recently proposed works focused on solving CRF using deep neural networks. Zheng et al. [33] formulated the mean-field updates of CRF as a recurrent neural network (RNN). Thus CRF parameters can be updated iteratively with back-propagation. There are mainly two drawbacks in the aforementioned works. First, most of them exploit CRF as a post-processing component to refine the segmentation labels, and hence cannot regulate the feature learning procedure. Second, most of these works allow CRF to smooth the segmentation map by encouraging spatial coherence [34]. Different from these works, our proposed CGA-CRF mainly contributes to feature aggregation, jointly trained with the network backbone. Moreover, our proposed CGA-CRF considers the contextual information extracted from both the convolution/interaction spaces and the contextual attributes can be generated from a learned attention mechanism.

Graph Neural Network in Medical Image Analysis. In recent years, graph neural network attracts the attention of researchers in medical image analysis for relational information learning [35]. Parisot et al. [36] constructed a population graph for degenerative disease classification where each node represents features from an individual patient. Li et al. [37] proposed a topology-adaptive graph neural network for landmark detection with applications on X-ray images. Chao et al. [38] introduced a graph neural network to model the relationship between inter lymph nodes for gross tumor detection. Other research works also applied graph neural network directly onto structured data. For example, Chen et al. [39] applied graph neural network for intracranial artery labeling using cerebral artery map data. Li et al. [40] fed the brain functional graph data into graph neural network for brain biomarker analysis. The aforementioned research works treated pre-localised regions of the image as graph vertices, which greatly limits the generalisation of these methods over different datasets. Our feature interaction graph treats the feature instances as vertices, which provides the relational learning ability and data independent transferability.

III. Proposed Method

In this section, we describe our proposed CANet for voxel-wise brain glioma segmentation. We first describe the proposed feature interaction graph in detail. Then we introduce the novel feature fusion module, CGA-CRF, which selectively aggregates the features generated from different contexts and learns to render optimal features. Finally, the formulation of the mean-field updates in CGA-CRF as sequential convolutional operations is described, enabling the network to achieve end-to-end training. The proposed segmentation framework is illustrated in Fig. 2. Supplementary B summarises the training steps of our proposed CANet.

Different from previous works, our proposed CANet can implicitly capture long-range relational information by reasoning on the feature interaction graph, which have not been fully studied in the literature. Both contexts (feature interaction graph and convolution) utilise the intermediate feature map $X \in \mathbb{R}^{N\times C}$ derived from the shared encoder backbone as input, where $N = H \times W \times D$ is the total number of the feature instances in the intermediate feature map. C is the number of the feature dimension. The graph context generates representations in the feature interaction graph space $X^G \in \mathbb{R}^{N\times C}$ and the convolution generates a coordinate space representation $X^C \in \mathbb{R}^{N\times C}$.

The main concept behind the design of CGA-CRF is to generate an optimal segmentation map $H \in \mathcal{H}$ associated with an MRI image $I \in \mathcal{I}$ by exploiting the relationship between the final representation $X^F \in \mathbb{R}^{N\times C}$ and the intermediate feature representation X with auxiliary long-range relational information X^G, generated from the interaction space with its convolution features X^C. Different from direct concatenation $X^F = \textit{concat}(X, X^G, X^C)$ or element-wise summation $X^F = X + X^G + X^C$, we aim to learn a set of latent feature representations X^F through a novel conditional random field. Since X^C and X^G may contribute differently during the learning X^F, we adopt the idea of an attention mechanism and generalise it to a gate node of CRFs. The gate node can regulate the information flow and discover the relevance between different contexts and latent features.

A. Context Guided Feature Extraction

1) Graph Context: **Projection with Adaptive Sampling** We first use the collected feature map to create a feature interaction space by constructing an interaction graph $G = \{V, E, A\}$, where V represents the set of nodes in the interaction graph, E represents the edges between the interaction nodes and A represents the adjacency matrix. Given a learned high dimensional feature $X = \{x_n\}_{n=1}^{N} \in \mathbb{R}^{N\times C}$ with $x_n \in \mathbb{R}^{1\times C}$ from the back-bone network, we first project the original feature onto the feature interaction space, generating a projected feature $X^{PROJ} = \{x_n^{proj}\}_{n=1}^{N} \in \mathbb{R}^{K\times C'}$. K is the number of the interaction nodes in the interaction graph and C' is the interaction space dimension. A naive method for

producing each element $x_n^{proj} \in X^{PROJ}, n = \{1, ..., K\}$ uses the linear combination of its neighbor elements [41]:

$$x_n^{proj} = \sum_{\forall m \in \mathcal{N}_n} w_{nm} x_m A[n, m] \tag{1}$$

where $\mathcal{N}_n$ denotes the neighbors of voxel n. The naive approach normally employs a fully-connected graph with redundant connections and parameters between the interaction nodes, being very difficult to optimise. More importantly, the linear combination method lacks an ability to perform adaptive sampling because different images contain different contextual information of brain glioma (e.g. location, size and shape). We deal with this issue by adopting the adaptive sampling strategy [42]:

$$\begin{aligned} \triangle m &= w_{n,m} x_n + b_{n,m} \\ x_n^{proj} &= \sum_{\forall m \in \mathcal{N}_n} w_{nm} \rho(x_m | \mathcal{V}, m, \triangle m) A[n, m] \end{aligned} \tag{2}$$

where $w_{n,m} \in \mathbb{R}^{3 \times (K \times C)}$ and $b_{n,m} \in \mathbb{R}^{3 \times 1}$ are the shift distances which are learned individually for each raw feature x_n through stochastic gradient decent. $\rho(\dot{})$ is the trilinear interpolation sampler which samples a shifted feature node around feature node x_m, given the learned deformation $\triangle m$ and the total set of interaction graph nodes $\mathcal{V}$.

Interaction Graph Reasoning After projected the input features onto the interaction graph G with K feature nodes $V = \{v_1, ..., v_k\}$ and edges E, we follow the definition of the graph convolution network [43], [44]. In particular, we define A^G as the graph adjacency matrix on $K \times K$ nodes and $W^G \in \mathbb{R}^{D \times D}$ as the weight matrix, and the formulation of the graph convolution operation is formulated as follows:

$$\begin{aligned} X^G &= \sigma(A^G X^{PROJ} W^G) \\ &= \sigma((I - \hat{A^G}) X^{PROJ} W^G) \end{aligned} \tag{3}$$

where $\sigma()$ is sigmoid activation function. We first apply Laplacian smoothing and update the adjacency matrix to $(I - \hat{A^G})$ so as to propagate the node feature over the entire graph. In practice, we implement $\hat{A^G}$ and W^G using a 1×1 convolution layer. We also implement I as a residual connection which maximises the gradient flow [45].

Re-Projection Once the reasoning has been finished, we re-project the features back to the original coordinate space with output $X^G \in \mathbb{R}^{N \times D}$. We use trilinear interpolation here to calculate each graph feature instance $x_n^g \in X^G, n \in \{1, ..., N\}$ after having transformed the features from the interaction space to the coordinate space. As a result, we have the interaction graph feature X^G with dimension D over N feature instances, identical to X^C.

2) Convolution Context Branch: The convolution context branch is composed of an encoder and a decoder with skip connections between these two components. The encoder reduces the spatial dimensionality of the feature map whilst the expansive path recovers the feature map's spatial dimensionality and the details of objects. One of the advantages of using this architecture is that it fully exploits the features with different scales of contextual information, where large scale features can be used to localise objects and small scale yet high dimensionality features can provide more detailed and accurate information for classification.

However, networks with 3D kernels contain more parameters to learn during feature extraction. It has been observed that training such 3D model often fails in various reasons such as over-fitting and gradient vanishing or exploding. In order to address the issues mentioned above, we deploy a deep supervised mechanism for better training the convolution context branch [46]. The proposed deep supervision mechanism thus reinforces the gradient flow and improves the discriminative capability during the training procedure.

Specifically, we use additional upsampling layers to reshape the features created at the deep supervised layer with the resolution of the final output. For each transform layer, we apply the softmax function to obtain additional dense segmentation maps. For these additional segmentation results, we calculate the segmentation errors with regards to the ground-truth segmentation maps. The auxiliary losses are combined with the loss from the output layer of the whole network and we further back-propagate the gradient for parameter updating during each iteration in the training stage.

We denote the set of the parameters in the deep supervised layers as $W^S = \{w^s\}_{s=1}^S$ and w^s as the parameters of the upsampling layer s. The auxiliary loss for a deep supervision layer s is formulated using cross-entropy:

$$L^s = \sum_{n=1}^{N} - \log \mathbb{1}(p(y_n | o_n^s; w^s)) \tag{4}$$

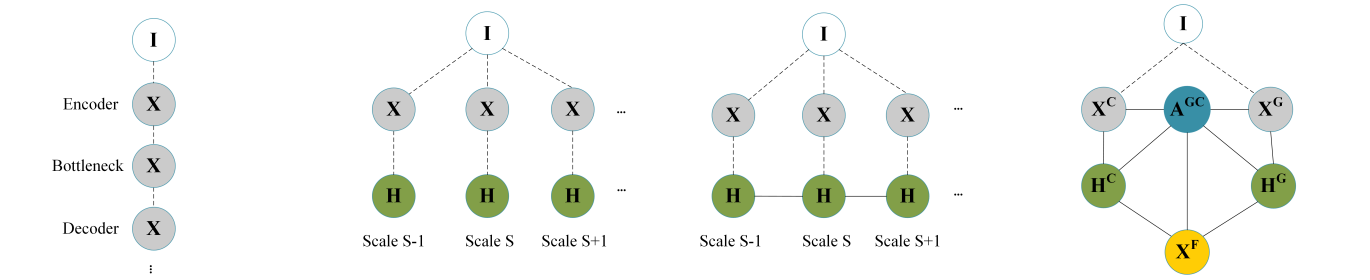
where $\mathbb{1}$ is the indicator function which is 1 if the segmentation result is correct, otherwise 0. $Y = \{y_n\}_{n=1}^N$ is the ground-truth of voxel n and $O = \{o_n\}_{n=1}^N$ is the predicted segmentation label of voxel n generated from the upsampling layer s. Finally, the deep supervision loss L_s can be integrated with the loss L^T from the final output layer. The parameters of the deep supervised layers W^S can be updated with the rest parameters W^{T-S} from the whole framework simultaneously using back-propagation:

$$\begin{aligned} L = L^T(Y|O; W^{T-S}, W^S) + \sum_{s=1}^{S} \delta^s L^s(O; w^s) \\ + \lambda(||W^{T-S}||^2 + \sum_{s=1}^{S}(||w^s||^2)) \end{aligned} \tag{5}$$

where δ^s represents the weight for the supervision loss of each upsampling layer. As the training procedure continues to approach to the optimal parameter sets, δ^s reduces gradually. The final operation of Eq. (5) is the $L2$-regularisation of the total trainable weights with scalar λ.

B. Context Guided Attentive CRF Fusion Module

We further propose a novel context guided attentive CRF module to perform feature fusion, motivated from two perspectives. The graph model of our proposed CGA-CRF is illustrated in Fig. 3. There are two reasons to use CGA-CRF for feature fusion. Firstly, assigning segmentation labels by maximising probabilities may result in incorrect boundary segmentation due to the neighboring voxels of sharing similar



(a) Encoder-Decoder Neural Network (b) Multi-Scale Neural Network (c) Multi-Scale CRFs (d) Context Guided-Attentive Fusion CRF

Fig. 3. Graph model illustration of previous feature fusion schemes: (a) Basic encoder-decoder neural network, (b) multi-scale neural network, (c) multi-scale CRF, and (d) our proposed context guided attentive CRF. I denotes the input 3D MRI image. S denotes a particular feature scale. X^C and X^G represent the hidden features generated from the convolutional operation and graph convolutional practice respectively. A^{GC} indicates the attention map generated from the corresponding feature X^C and X^G. Best viewed in color.

feature representation. Secondly, previous works fuse features from different sources by using a channel-wise concatenation or element-wise summation mechanism. However, these mechanisms simplifies the relationship between different source feature maps, which may result in information loss. Different from previous related works and using the inference ability of a probabilistic graphical model, we employ the conditional random field model to learn optimised latent fusion features for final settlement. As information from different contexts may contribute to the final results with different degrees, we integrate the attention gates of the CGA-CRF to regulate how much information should flow between features. We further show the implementation of CGA-CRF mean-field updates with sequential convolution operations, which allows our CGA-CRF fusion module can be integrated with any neural networks as sequential layers and trained in an end-to-end fashion. Compared with previous architectures such as encoder-decoder neural network (Fig. 3 (a)) and multi-scale neural network (Fig. 3 (b)), our proposed CGA-CRF (Fig. 3 (d)) has a strong inference ability and can jointly learn the hidden representation of features encoded by the neural network backbone, improving the generalisation ability of the segmentation model. Compared with previous architectures such as multi-scale CRF (Fig. 3 (c)), our proposed CGA-CRF model first uses an attention gate by directly modeling the cost energy in the network (Eq. (7)). The attention gate thus regulates the information flow from the features encoded by the backbone neural network to the latent representations by minimising the total energy cost. We evaluate the effectiveness of each component in the experiment section.

1) Definition: Given the feature map $X^C = \{x_n^c\}_{n=1}^N$ from the convolution context branch and the feature map $X^G = \{x_n^g\}_{n=1}^N$ from the interaction graph branch, our goal is to estimate the relationship between the hidden representation $H^G = \{h_n^g\}_{n=1}^N, H^C = \{h_n^c\}_{n=1}^N$, the attention variable $A^{GC} = \{a_n^{gc}\}_{n=1}^N$ and the final fused representation $X^F = \{x_n^f\}_{n=1}^N$. We formalise the problem by designing CGA-CRF with a gibbs distribution:

$$P(X^F, A^{GC}|I,\Theta) = \frac{1}{Z(I,\Theta)} exp\{-E(X^F, A^{GC}, I, \Theta)\} \tag{6}$$

where $E(X^F, A^{GC}, I, \Theta)$ is the associated energy:

$$\begin{aligned} E(X^F, A^{GC}, I, \Theta) =& \Phi^G(H^G, X^G) + \Phi^C(H^C, X^C) \\ &+ \Psi^{GC}(H^G, H^C, A^{GC}) \end{aligned} \tag{7}$$

where I is the input 3D MRI image and Θ is the set of parameters. In Eq. (7), Φ^G is the unary potential between the latent graph representation H^G and the graph features X^G. Φ^C is the unary potential related to latent convolution representation H^C and convolution feature X^C. In order to drive the estimated latent representation H towards the observation X, we use the Gaussian function created in previous works [47]:

$$\Phi(H, X) = \sum_{n=1}^{N} \phi(h_n, x_n) = -\sum_{n=1}^{N} \frac{1}{2}||h_n - x_n||^2. \tag{8}$$

The final term shown in Eq. (7) is the attention guided pairwise potential between the latent convolution representation H^C and the latent graph representation H^G. The attention term A^{GC} controls the information flow between the two latent representations where the graph representation may or may not contribute to the estimated convolution representation. We define:

$$\begin{aligned} \Psi^{GC}(H^G, H^C, A^{GC}) &= \sum_{n=1}^{N} \sum_{m \in N_n} \psi(a_m^{gc}, h_n^c, h_m^g) \\ &= \sum_{n=1}^{N} \sum_{m \in N_n} a_m^{gc} h_n^c \Upsilon_{n,m}^{GC} h_m^g \end{aligned} \tag{9}$$

The $\Upsilon_{n,m}^{GC} \in \mathbb{R}^{D^G \times D^C}$ is the kernel potential associated with hidden feature maps H^G and H^C, where D^G, D^C represents the dimensionality of the features X^G and X^C respectively.

2) Inference: By learning latent feature representations to minimise the total segmentation energy, the system can produce an appropriate segmentation map, *e.g.* maximum a posterior $P(X^F, A^{GC}|I, \Theta)$. However, the optimisation of $P(X^F, A^{GC}|I, \Theta)$ is intractable due to the computational complexity of the normalisation constant $Z(I, \Theta)$, which is exponentially proportional to the cardinality of X^F and A^{GC}. Therefore, in order to derive the maximum a posterior in an

efficient way, we adopt mean-field updates to approximate a complex posterior probability distribution:

$$P(X^F, A^{GC}|I,\Theta) \approx Q(X^F, A^{GC}) \approx Q(H^G, H^C, A^{GC}) = \prod_{n=1}^{N} q_n(h_n^G) q_n(h_n^C) q_n(a_n^{gc}) \tag{10}$$

Here, we use the product of independent marginal distributions $q(h^g)$, $q(h^c)$ and $q(a^{gc})$ to approximate the complex distribution $P(X^F, A^{GC}, I, \Theta)$. To achieve a satisfactory approximation, we minimise the Kullback-Leibler (KL) divergence $D_{KL}(Q||P)$ between the two distributions Q and P. By replacing the definition of the energy $E(X^F, A^{GC}, I, \Theta)$, we formulate the KL divergence in Eq. (10) as follows:

$$D_{KL}(Q||P) = \sum_h Q(h) \ln(\frac{Q(h)}{P(h)}) = \sum_h Q(h)E(h) + \sum_h Q(h) \ln Q(h) + \ln Z \tag{11}$$

From Eq. (11), we minimise the KL divergence by directly minimising the free energy $FE(Q) = \sum_h Q(h)E(h) + \sum_h Q(h)\ln(Q(h))$. In $FE(Q)$, the first item represents the cost for labelling each voxel and the second item represents the entropy of distribution Q. We can further expand the expression of $FE(Q)$ by replacing Q and E with Eqs. (10) and (7) respectively:

$$FE(Q) = \sum_{n=1}^{N} q_n(h_n^g) q_n(h_n^c) q_n(a_n^{gc})(\Phi^G + \Phi^C + \Psi^{GC}) + \sum_{n=1}^{N} q_n(h_n^g) q_n(h_n^c) q_n(a_n^{gc})(\ln(q_n(h_n^g) q_n(h_n^c) q_n(a_n^{gc}))) \tag{12}$$

Eq. (12) shows that the problem of minimising $FE(Q)$ can be transferred to a constrained optimisation problem with multiple variables, formulated below:

$$\min_{q_n(h_n^g), q_n(h_n^c), q_n(a_n^{gc})} FE(Q), \forall n \in N \quad \text{s.t.} \sum_{n=1}^{N} q_n(h_n^g) = 1, \sum_{n=1}^{N} q_n(h_n^c) = 1, \int_0^1 q_n(a_n^{gc}) da_n^{gc} = 1 \tag{13}$$

We can calculate the first order partial derivative by differentiating $FE(Q)$ w.r.t each variable. For example, we have:

$$\frac{\partial FE}{\partial q_n(h_n^c)} = \phi^c(h_n^c, x_n^c) + \sum_{m \in \mathcal{N}_n} \mathbb{E}_{q_m(a_m^{gc})}\{a_m^{gc}\} \mathbb{E}_{q_m(h_n^g)} \psi^{gc}(h_n^c, h_m^g) - \ln q_n(h_n^c) + \text{const} \tag{14}$$

By assigning 0 to the left hand side of Eq. (14), we reach:

$$q_n(h_n^c) \propto \exp\{\phi^c(h_n^c, x_n^c) + \sum_{m \in \mathcal{N}_n} \mathbb{E}_{q_m(a_m^{gc})}\{a_m^{gc}\} \mathbb{E}_{q_m(h_m^g)} \psi(h_n^c, h_m^g)\} \tag{15}$$

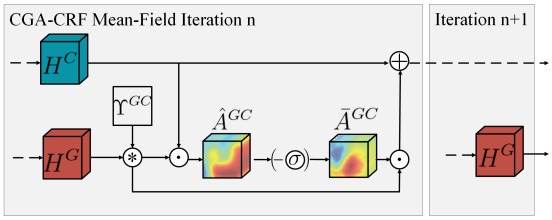


Fig. 4. Details of the mean-field updates within CGA-CRF. The circled symbols indicate message-passing operations within the CGA-CRF block. Best viewed in colors.

Eq. (15) shows that, once the other two independent variables $q(h^g)$ and $q(a^{gc})$ are fixed, how $q(h^c)$ is updated during the mean-field approximation. Furthermore, we follow the above procedure and obtain the updating of the remaining two variable as follows:

$$q_n(h_n^g) \propto \exp\{\phi^g(h_n^g, x_n^c) + \mathbb{E}_{q_m(a_m^{gc})}\{a_m^{gc}\} \sum_{m \in \mathcal{N}_i} \mathbb{E}_{q_m(h_m^c)} \psi(h_n^c, h_m^g)\} \tag{16}$$

$$q_n(a_n^{gc}) \propto \exp\{a_n^{gc} \mathbb{E}_{q_n(h_n^c)}\{\sum_{m \in \mathcal{N}_n} \mathbb{E}_{q_m(h_m^g)}\{\psi(h_n^c, h_m^g)\}\}\} \tag{17}$$

where $\mathbb{E}_{q()}$ represents the expectation with respect to the distribution $q()$. Eqs. (15-17) shown above denote the computational procedure of seeking an optimal posterior distributions of h^c, h^g and a^{gc} during the mean-field approximation. Intuitively, Eq. (15) shows that, the latent convolution feature h_n^c for voxel n can be used to describe the observation, referred to feature x_n^c. Afterwards, we use the re-weighted messages from the latent features of the neighboring voxels to learn the co-occurrent relationship of the pixels. The attention balance between the latent convolution and the graph features for voxel n allows us to re-weight the pairwise potential message from the neighbours of voxel n, and then use the attention variable to re-weight the total value of voxel n. By denoting $\bar{a}_n^{gc} = \mathbb{E}_{q(a_n^{gc})}\{a_n^{gc}\}$ and $\bar{h}_n = \mathbb{E}_{q(h_n)}\{h_n\}$, we have the feature update as follows:

$$\bar{h}_n^g = x_n^g + \bar{a}_n^{gc} \sum_{m \in \mathcal{N}_n} \Upsilon_{n,m}^{GC} \bar{h}_m^c \tag{18}$$

$$\bar{h}_n^c = x_n^c + \sum_{m \in \mathcal{N}_n} \bar{a}_m^{gc} \Upsilon_{n,m}^{GC} \bar{h}_m^g \tag{19}$$

$\bar{a}_n^{gc}$ is also derived from the probabilistic distribution, *i.e.* its value lies in $[0, 1]$. Here, we choose the Sigmoid function to formulate the updates for $\bar{a}_n^{gc}$:

$$\bar{a}_n^{gc} = \sigma(-\sum_{m \in \mathcal{N}_n} a_m^{gc} h_n^c \Upsilon_{n,m}^{GC} h_m^g) \tag{20}$$

where $\sigma(.)$ denotes the sigmoid activation function.

3) CGA-CRF Inference as Convolutional Operations: We implement the mean-field updates of CGA-CRF as sequential convolutional operations in order that the CGA-CRF can be trained with any neural network in an end-to-end fashion. The algorithm for implementing mean-field approximation using convolutional operations is described in Algorithm 1.

Algorithm 1: Algorithm for Mean-Field Approximation of CGA-CRF.

Input: Tensor instance from the feature interaction graph output X^G and convolution output X^C. Initialise hidden graph feature map instance H^G with X^G. Initialise hidden convolutional feature map instance X^C with X^C

Output: Estimated optimised feature map H^C as X^F.

1: **while** in iteration number **do**
2: $\hat{A}^{GC} \leftarrow H^C \odot (\Upsilon^{GC} * H^G)$;
3: $\bar{A}^{GC} \leftarrow \sigma(-(\hat{A}^{GC}))$;
4: $H^G \leftarrow \Upsilon^{GC} * H^G$;
5: $\bar{H}^C \leftarrow \bar{A}^{GC} \odot H^G$;
6: $H^C \leftarrow X^C \oplus \bar{H}^C$;
7: **end while**
8: **return** Optimised feature map H^C as X^F.

Following Algorithm 1, we compile each iteration of the mean-field updates by a set of convolution operations. The output of the previous sequential convolution blocks are sent to the next sequential convolution blocks to complete one iteration. A graph illustration of Algorithm 1 is shown in Fig. 4. Through this implementation, we aim to jointly estimate the hidden feature maps H^G, H^C and the attention map A^{GC} based on the derivation shown in Section III-B2 in a data-driven manner.

Based on Eq. (20), we implement the update process of attention map A^{GC} as follows: (1) Execute the message passing between hidden feature maps H^G and H^C, $\hat{A}^{GC} \leftarrow H^C \odot (\Upsilon^{GC} * H^G)$. Here, Υ^{GC} is a convolution kernel sliding on H^G. We set the kernel size of Υ^{GC} as 3. $*$ and $\odot$ represents the convolutional and element-wise product respectively. (2) Normalise the $\hat{A}^{GC}$ using sigmoid function $\bar{A}^{GC} \leftarrow \sigma(-(\hat{A}^{GC}))$.

We utilise the attention map A^{GC} as a switch to regulate the message flow when updating H^G and H^C. Specifically, the mean-filed update of H^G and H^C can be implemented as follows: (1) Execute the message passing on H^G: $H^G \leftarrow \Upsilon^{GC} * H^G$. (2) Multiply the interaction graph feature with attention map $\bar{H}^C \leftarrow \bar{A}^{GC} \odot H^G$. (3) Update H^C by adding the unary potential using residual connections: $H^C \leftarrow X^C \oplus \bar{H}^C$, where $\oplus$ represents the element-wise summation.

By implementing the mean-field updates as sequential convolution operations, H^G, H^C and A^{GC} can be learnt and updated jointly. Note that the generalised mean field approximation is guaranteed to converge to a local optimum rather than a global optimum. Thus, to reduce the computational time of our proposed CGA-CRF, we have examined the iteration numbers of mean-field approximation of our CGA-CRF progressively and set the iteration number as 5 to reach a trade-off between competitive performance and small parameter size.

IV. Experimental Setup

To demonstrate the effectiveness of the proposed CANet for brain glioma segmentation, we conduct experiments on three publicly available datasets: the Multimodal Brain Tumor Segmentation Challenge 2017 (BraTS2017), the Multimodal Brain Tumor Segmentation Challenge 2018 (BraTS2018) and the Multimodal Brain Tumor Segmentation Challenge 2019 (BraTS2019) [48]–[50]. Supplementary C presents data augmentation and implementation settings.

Datasets. The **BraTS2017**[1] consists of 285 cases of patients in the training set and 44 cases in the validation set. **BraTS2018**[2] shares the same training set with BraTS2017 and includes 66 cases in the validation set. **BraTS2019**[3] expands the training set to 335 cases and the validation set to 125 cases. Each case is composed of four MR sequences, namely native T1-weighted (T1), post-contrast T1-weighted (T1ce), T2-weighted (T2) and Fluid Attenuated Inversion Recovery (FLAIR). Each sequence has a 3D MRI volume of 240×240×155. Ground-truth annotation is only provided in the training set, which contains the background and healthy tissues (label 0), necrotic and non-enhancing tumor (label 1), peritumoral edema (label 2) and GD-enhancing tumor (label 4). We first consider the 5-fold cross-validation on the training set where each fold contains (by random division) 228 cases for training and 57 cases for validation. We then evaluate the performance of the proposed method on the validation set. The validation result is generated from the official server of the contest to determine the segmentation accuracy of the proposed methods.

Evaluation Metrics. Following previous works [28], [23], [49], the segmentation accuracy is measured by Dice score, Sensitivity, Specificity and Hausdorff95 distance respectively. In particular,

- Dice score: $Dice(P,T) = \frac{|P_1 \cap T_1|}{(|P_1|+|T_1|)/2}$
- Sensitivity: $Sens(P,T) = \frac{|P_1 \cap T_1|}{|T_1|}$
- Specificity: $Spec(P,T) = \frac{|P_0 \cap T_0|}{|T_0|}$
- Hausdorff Distance: $Haus(P,T) = max\{sup_{p \in P_1} inf_{t \in T_1} d(p,t), sup_{t \in T_1} inf_{p \in P_1} d(t,p)\}$

where P represents the model prediction and T represents the ground-truth annotation. T_1 and T_0 are the subset voxels predicted as positives and negatives for the tumor regions. Similar set-ups are made for P_1 and P_0. Furthermore, the Hausdorff95 measures the distance when comparing model prediction against ground-truth segmentation [51]. sup represents the supremum and inf represents the infimum. For each metric, three regions namely enhancing tumor (ET, label 1), whole tumor (WT, labels 1, 2 and 4) and the tumor core (TC, labels 1 and 4) are evaluated individually.

V. Results and Discussion

In this section, we present both quantitative and qualitative experimental results of different methods. We first conduct an ablation study of our method to show the effective impact of building a feature interaction graph and CGA-CRF on the segmentation performance. We also perform extensive analysis on the encoder backbone and different iteration numbers of approximation in CGA-CRF. Afterwards, we compare our approach with several State-of-The-Art methods on different datasets. Finally, we present the analysis of failure cases.

[1] https://www.med.upenn.edu/sbia/brats2017.html
[2] https://www.med.upenn.edu/sbia/brats2018.html
[3] https://www.med.upenn.edu/cbica/brats-2019/

TABLE I

QUANTITATIVE RESULTS OF THE CANET COMPONENTS BY FIVE FOLD CROSS-VALIDATION FOR THE BRATS2017 TRAINING SET (DICE, SENSITIVITY AND SPECIFICITY). ALL THE METHODS ARE BASED ON CANET WITH UNET AS THE BACKBONE. THE BEST RESULT IS SHOWN IN BOLD TEXT AND THE RUNNER-UP RESULT IS UNDERLINED.

	DICE			Sensitivity			Specificity			Hausdorff95		
Backbone+	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC
CC	**0.686**	0.875	0.821	0.857	0.925	0.863	0.997	0.991	0.996	**6.791**	6.886	7.939
GC	0.637	0.894	0.822	**0.977**	**0.970**	**0.944**	0.987	0.987	**0.997**	9.899	6.403	5.812
CC+GC+Concatenation	0.682	0.861	0.803	0.857	0.922	0.861	0.997	0.989	0.994	7.755	9.377	11.432
CC+GC+CGA-CRF	0.685	**0.903**	**0.873**	0.807	0.924	0.870	**0.997**	**0.993**	0.996	7.804	**3.569**	**4.036**

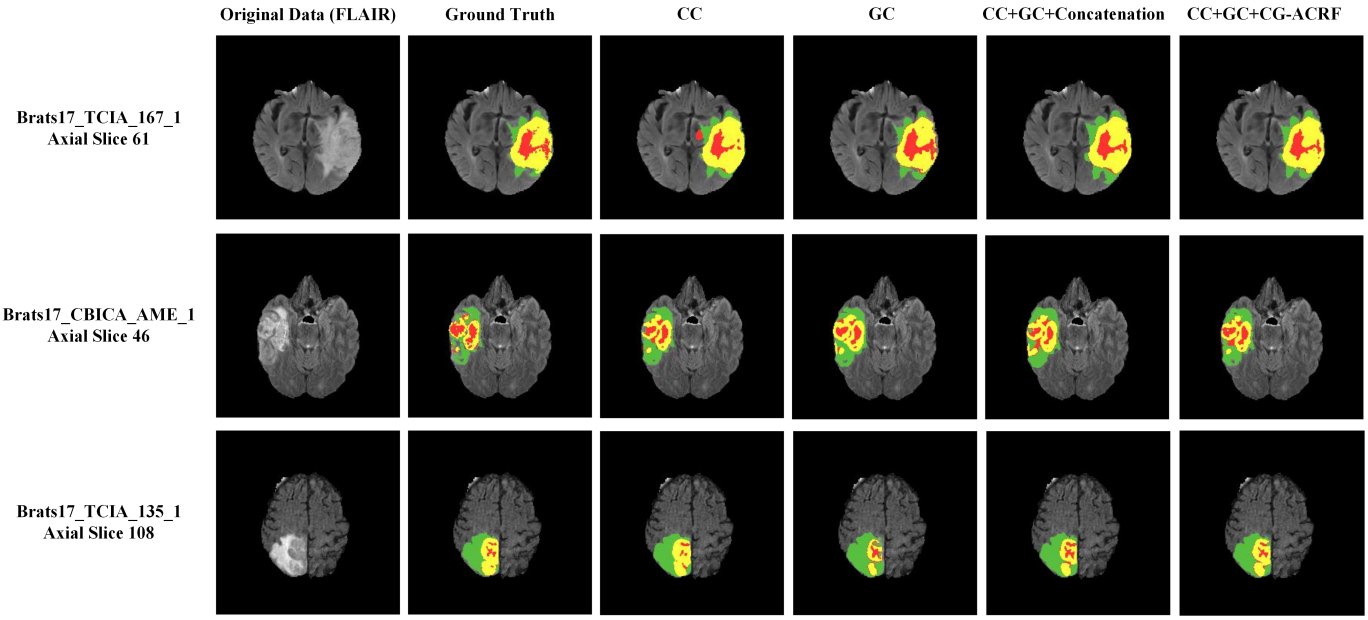


Fig. 5. Qualitative comparison of different baseline models and the proposed CANet by cross validation on the BraTS2017 training set. From left to right, each column represents the input FLAIR data, ground truth annotation, segmentation result of CANet with only the convolution branch, segmentation result of CANet with only the graph convolution branch, segmentation output of CANet with HCA-FE and concatenation fusion scheme, segmentation output of CANet with HCA-FE and CG-ACRF fusion module. Best viewed in colors.

A. Ablation Studies

We first evaluate the impact of building the feature interaction graph and CGA-CRF. To this end, we apply 5-fold cross-evaluation on the BraTS2017 training set and report the mean results. Table I shows the quantitative results, while the qualitative results can be found in Fig. 5 as an example of the segmentation outputs. We start from two baselines. The first baseline is the fully convolution network with deep supervision on the backbone convolution encoder (CC). The second baseline only uses the proposed interaction graph in the convolution encoder without deep supervision (GC). We then evaluate the proposed (whole) CANet system (CC+GC) with concatenated feature maps from CC and GC together without any additional feature fusion method. Finally, we evaluate the proposed feature fusion module CGA-CRF, which takes the feature map with different contexts and outputs the optimal latent feature map for the final segmentation. For the experiments shown in Table I and Fig. 5, we use the encoder of UNet as the backbone network with 5 iterations in our CGA-CRF. The experiments described later include the analysis on different backbones with iteration numbers.

From Table I, we observe that the GC obtains better performance than CC. For dice scores, GC achieves 0.894 for the entire tumor and 0.822 for the tumor core. CC only achieves a dice score of 0.875 on the entire tumor and 0.821 on the tumor core, which is 2% and 0.2% lower than those by GC respectively. For hausdorff95, GC achieves 6.403 on the entire tumor and 5.812 on the tumor core. CC achieves 6.886 and 7.939, which are 0.493 and 2.127 higher than those of GC on the entire tumor and the tumor core, respectively. From Fig. 5, we observe that GC can accurately predict individual regions. For example, the GD-enhanced tumor region normally does not appear at the outside of the tumor region. This superior performance may benefit from the information learned from the feature interactive graph as the feature nodes of different tumor regions have strong structural association between them. Learning the relationship may help the system to predict correct labels of the tumor regions. However, the sensitivity of GC is much higher than that of CC. In Table I, for example, the sensitivity score of GC is higher than that of CC: 12.02% higher on the enhancing tumor, 4.469% higher on the entire tumor, 8.104% higher on tumor core, respectively. We observe poor segmentation results at the NCR/ECT region by GC, inferior to CC with the ground truth shown in Fig. 5.

We then evaluate the complete CANet with the extracted feature maps by CC and GC simultaneously. Here, we fuse

TABLE II
QUANTITATIVE RESULTS FOR DIFFERENT ITERATION NUMBERS BY CG-ACRF MEAN-FIELD APPROXIMATION ON THE FIVE FOLD CROSS-VALIDATION OF THE BRATS2017 TRAINING SET WITH RESPECT TO DICE, SENSITIVITY, SPECIFICITY AND HAUSDORFF95. THE BEST RESULT IS IN BOLD AND THE RUNNER-UP RESULT IS UNDERLINED.

	Dice			Sensitivity			Specificity			Hausdorff95		
Iteration #	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC
1	0.657	0.861	0.790	**0.901**	0.920	0.852	0.995	0.990	0.994	7.997	7.749	10.488
3	0.681	0.873	0.807	0.873	0.923	0.869	0.996	0.990	0.994	**7.614**	6.801	8.941
5	**0.685**	**0.903**	**0.873**	0.807	0.924	**0.870**	**0.997**	**0.993**	**0.996**	7.804	**3.569**	**4.036**
7	0.664	0.855	0.769	0.854	0.921	0.860	0.996	0.990	0.993	9.850	9.720	12.042
10	0.685	0.850	0.784	0.837	**0.931**	0.858	0.997	0.988	0.993	8.067	11.149	11.650

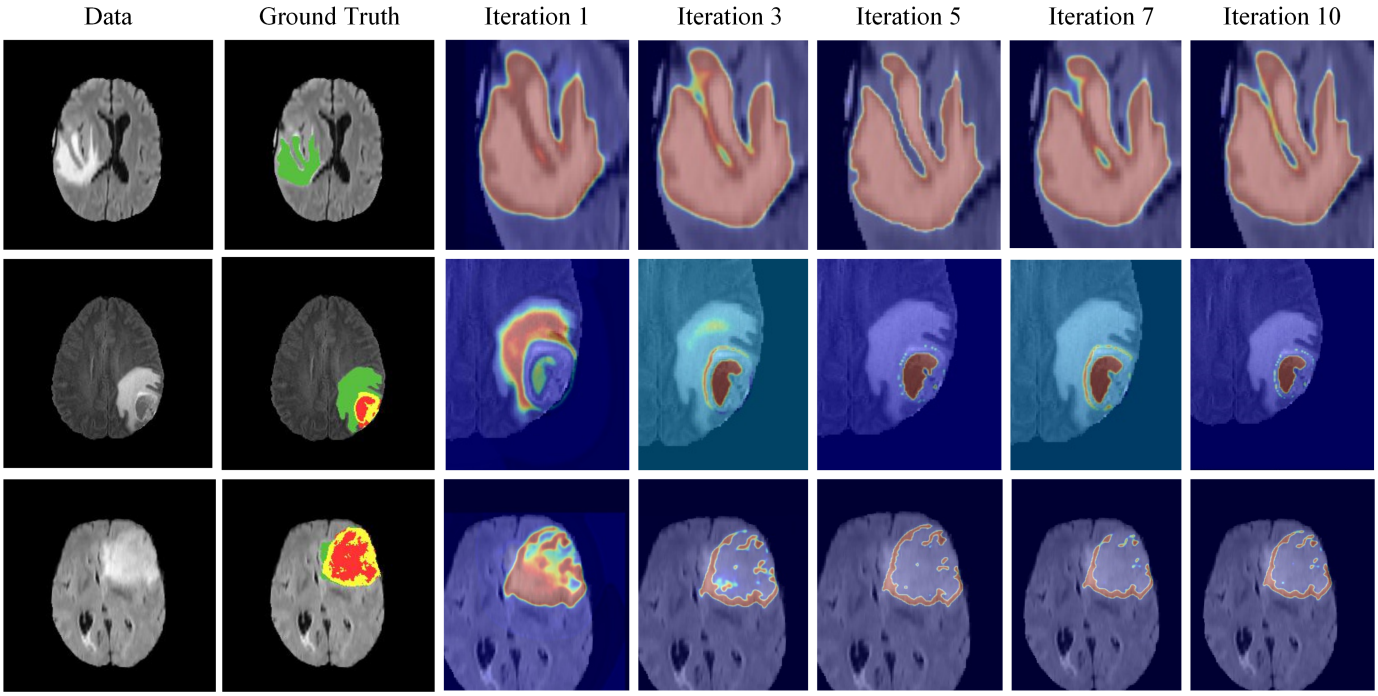


Fig. 6. Examples to illustrate the results with different iteration numbers by mean-field approximation in the proposed CG-ACRF. Columns from top to bottom represent different patient cases. Rows from left to right indicate FLAIR data, ground truth annotation, attentive map generated by CANet with different iteration numbers (from 1 to 10) in CG-ACRF respectively. Best viewed in colors.

the feature maps of CC and GC using a naive concatenation method, which has less over-segmentation results. Depicted in Table I, the sensitivity of CC+GC is much lower than that of GC. The sensitivity of CC+GC is 0.857 on the enhancing tumor (ET), 0.922 on the whole tumor (WT) and 0.861 on the tumor core (TC), respectively. From Fig. 5, we witness that by introducing the feature interaction graph, the segmentation model can correct some misclassified regions produced by CC. However, the concatenation fusion method does not demonstrate any benefit on the overall segmentation performance. CC+GC has a dice score of 0.861 on the whole tumor and 0.803 on the tumor core, which are 3.292% and 1.94% lower than those of GC respectively. We also observe the loss of the boundary information shown in Fig. 5, especially the boundaries of NCR/ECT and GD-enhancing tumors excessively shrinks compared with those of GC and CC.

We finally evaluate the effectiveness of our proposed CGA-CRF. By introducing the CGA-CRF fusion module, our segmentation model outperforms the other methods. Benefiting from the inference ability of CGA-CRF, it presents a satisfactory segmentation output. For the whole tumor and the tumor core, its Dice scores are 0.903 and 0.873 respectively, which are the top scores in the leader-board. Its Hausdorff95 also is the lowest. For the whole tumor and the tumor core, its hausdorff95 values are 3.569 and 4.036 respectively. Referring to much lower sensitivity scores reported in Table I, we conclude that the superior performance has been achieved by the complete CANet. The same conclusion can be drawn from Fig. 5 where CGA-CRF can detect optimal feature maps that benefit the downstream deconvolution networks and outline small tumor cores and edges, which may be lost when we use a down-sampling operation in the encoder backbone.

Backbone Test We then evaluate the effectiveness of different encoder backbones. To do so, we use 5 fold cross-validation on the BraTS2017 training set with complete CC+GC and 5-iteration CGA-CRF. We here choose the State-of-The-Art encoder backbones, e.g. VGG16, ResNet18, ResNet30, ResNet50 and UNet encoder path. For each backbone, we feed the feature map from the last convolution block into the feature interaction graph branch to extract the

TABLE III
QUANTITATIVE RESULTS OF THE STATE-OF-THE-ART MODELS BY CROSS-VALIDATION ON THE BRATS2017 TRAINING SET WITH RESPECT TO DICE, SENSITIVITY, SPECIFICITY AND HAUSDORFF. THE BEST RESULT IS SHOWN IN BOLD AND THE RUNNER-UP RESULT IS UNDERLINED.

	Dice			Sensitivity			Specificity			Hausdorff95		
Model	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC
3D-UNet [52]	0.706	0.865	0.810	0.803	0.906	0.829	0.998	0.990	0.995	6.624	8.193	8.958
No-New Net [53]	**0.741**	0.871	0.812	0.767	0.893	0.831	**0.999**	0.992	0.995	**3.930**	7.055	7.641
Attention UNet [54]	0.672	0.863	0.778	**0.847**	0.900	0.862	0.996	0.990	0.992	9.347	9.676	10.668
PRUNet [55]	0.710	0.891	0.814	0.788	0.900	0.841	0.998	0.990	0.996	7.205	7.414	9.187
3D-ESPNet [56]	0.690	0.895	0.844	0.805	**0.947**	**0.881**	0.997	0.990	0.997	6.894	4.156	5.778
CANet (Ours)	0.685	**0.903**	**0.873**	0.807	0.924	0.870	0.997	**0.993**	0.996	7.804	**3.569**	**4.036**

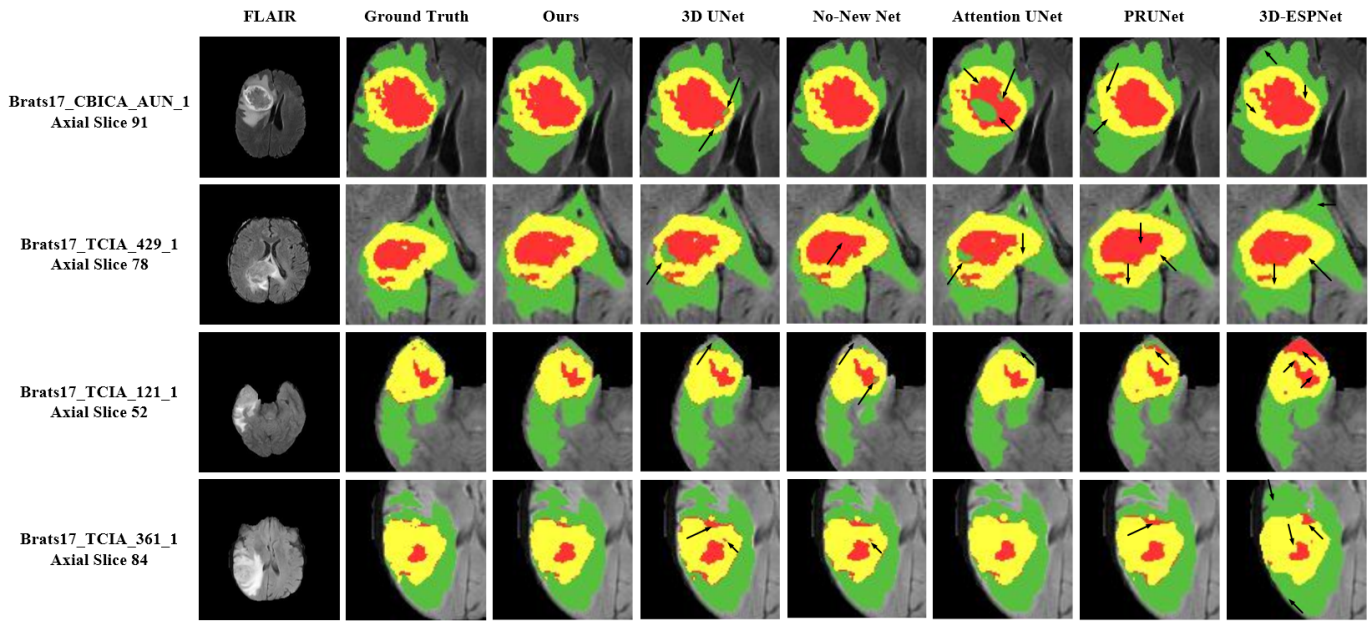


Fig. 7. Examples of segmentation results by cross validation on the BraTS2017 training set. Qualitative comparisons with other brain glioma segmentation methods are presented. The eight columns from left to right show the frames of the input FLAIR data, the ground truth annotation, the results generated from our CANet (UNet encoder backbone and 5-iteration CGA-CRF), 3DUNet [52], NoNewNet [53], Attention UNet [54], PRUNet [56], respectively. Black arrows indicate the failure cases in these comparison methods. Best viewed in colors.

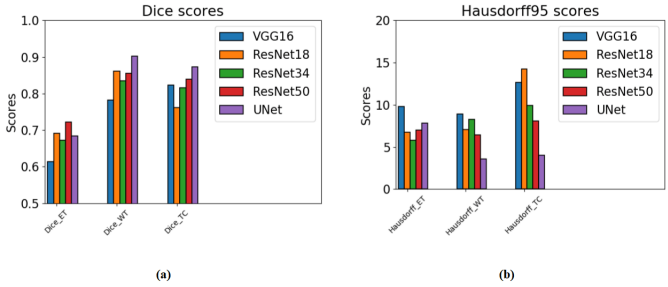


Fig. 8. Performance comparison with different encoder backbones: (a) and (b) indicate the comparisons with dice score and hausdorff95 by cross validation on the BraTS2017 training set using different encoder backbones respectively. Best viewed in colors.

interaction graph contexts and feed the feature map from the second last convolution block into the convolution branch to generate deep supervised feature maps. This practice has been proved to be effective, efficient and simple. The segmentation results with respect to Dice and Hausdorff95 are shown in Fig. 8. ResNet outperforms the VGG16 mainly due to the involved residual connection and batch normalisation. However, comparing ResNet and the encoder of UNet, the encoder of UNet achieves better segmentation performance in terms of Dice and Hausdorff95 due to the multiple scale feature maps and skip connection for feature fusion. We choose the encoder of UNet as the backbone network for the final segmentation model in our approach.

Iteration Test we manually set the iteration number in the mean-field approximation of CG-ACRF. Since the mean-field approximation can only guarantee a local optimal, we examine the consequence of different iteration numbers. Table II reports the quantitative result of using different iteration numbers, i.e. 1, 3, 5, 7, and 10. With increasing iterations, our proposed model performs better. However, we observe that no additional performance benefit can be gained when the iteration number exceeds 5. Fig. 6 presents the probability map during segmentation, where the light color represents the region with a lower probability while the dark color represents the area with a higher probability. We observe that using only one iteration, CANet can outline the region of interest using the fused feature maps. By increasing the iteration number to 3 or 5, CG-ACRF can gradually extract an optimal feature map, leading to accurate segmentation. We further increase the

TABLE IV
QUANTITATIVE COMPARISONS BETWEEN CANET AND THE OTHER STATE OF THE ART TECHNIQUES ON THE BRATS2017 VALIDATION SET WITH RESPECT TO DICE AND HAUSDORFF95. THE BEST RESULTS OUT OF EACH CATEGORY ARE SHOWN IN BOLD. '-' DEPICTS THAT THE RESULT OF THE ASSOCIATED METHOD HAS NOT BEEN REPORTED YET. ⋆ AND † REPRESENTS THE FINAL WINNER AND RUNNER-UP SOLUTION RESPECTIVELY.

			Dice			Hausdorff95			Model Parameter
Approach	Method		ET	WT	TC	ET	WT	TC	
Ensemble	Kamnitsas et al. [24] ⋆	Mean:	0.738	0.901	0.797	4.500	4.230	6.560	-
		StdDev:	-	-	-	-	-	-	
	Wang et al. [28] †	Mean:	**0.786**	**0.905**	**0.838**	**3.282**	**3.890**	**6.479**	5.95E5
		StdDev:	-	-	-	-	-	-	
	Zhao et al. [57]	Mean:	0.754	0.887	0.794	-	-	-	-
		StdDev:	-	-	-	-	-	-	
	Isensee et al. [58]	Mean:	0.732	0.896	0.797	4.550	6.970	9.480	-
		StdDev:	-	-	-	-	-	-	
	Jungo et al. [59]	Mean:	0.749	0.901	0.790	5.379	5.409	7.487	-
		StdDev:	0.277	0.086	0.239	10.068	9.710	8.935	
Single Prediction	Islam et al. [60]	Mean:	0.689	0.876	0.761	12.938	9.820	12.361	1.34E8
		StdDev:	0.304	0.086	0.221	26.453	13.516	20.826	
	Lopez et al. [61]	Mean:	0.567	0.783	0.685	23.828	30.316	38.077	1.57E7
		StdDev:	-	-	-	-	-	-	
	Shaikh et al. [62]	Mean:	0.650	0.870	0.680	-	-	-	2.31E7
		StdDev:	0.320	0.110	0.340	-	-	-	
	Castillo et al [63]	Mean:	0.690	0.860	0.690	-	-	-	2.22E7
		StdDev:	-	-	-	-	-	-	
	Li et al. [64]	Mean:	0.704	0.871	0.682	7.699	10.396	13.062	3.80E5
		StdDev:	0.307	0.083	0.304	14.407	15.754	17.573	
	Jesson et al. [65]	Mean:	0.713	**0.899**	0.751	6.980	**4.160**	**8.650**	4.29E6
		StdDev:	0.291	0.070	0.240	12.100	3.370	9.350	
	Roy et al. [66]	Mean:	0.716	0.892	0.793	6.612	6.735	9.806	-
		StdDev:	-	-	-	-	-	-	
	Pereira er al. [67]	Mean:	0.719	0.889	0.758	5.738	6.581	11.100	-
		StdDev:	-	-	-	-	-	-	
	CANet (Ours)	Mean:	**0.728**	0.892	**0.821**	**5.496**	7.392	10.122	3.34E7
		StdDev:	0.286	0.082	0.167	11.690	11.917	16.966	

iteration number to 7 and 10 but no further improvement has been made. Therefore, we set the iteration number to 5 as a working trade-off between the segmentation performance and the number of the engaged parameters.

B. Comparison with State-of-The-Art methods

We choose several State-of-The-Art deep learning based brain glioma segmentation methods, including 3D UNet [52], Attention UNet [54], PRUNet [55], NoNewNet [53] and 3D-ESPNet [56]. We first consider 5-fold cross-validation using the BraTS2017 training set. Each fold contains randomly chosen 228 cases for training and 57 cases for validation. In these cross-validation experiments, we consider CANet with CC+GC and CGA-CRF fusion modules with 5-iteration, leading to the best performance in the ablation tests. As shown in Table III, our CANet outperforms the other State-of-The-Art methods on several metrics while the results of the proposed method is competitive for the other metrics. The Dice score of CANet is 0.903 and 0.873 for the whole tumor and the tumor core respectively, where the former is 8% higher and the later is 3% higher than individual runner up results. The Hausdorff95 values of CANet are 3.569 and 4.036 for the whole tumor and the tumor core, which are much lower than the runner up scores, i.e. 4.156 and 5.778, respectively. Based on the individual score generated from the official evaluation server, we argue that the Dice score of ET from our proposed CANet is affected by the data-imbalance issue. The evaluation of enhancing tumor only considers the prediction of pertumoral edema, which only exists in the High-Grade Glioma (HGG) patients. As the training set contains more HGG cases than the LGG (Lower-Grade Glioma) cases, our system may learn a bias and make inaccurate prediction on some LGG cases in validation as the HGG cases contain false positives for the prediction of pertumoral edema labels. This false positive prediction leads to 0 Dice score on ET instead of 1 for LGG cases, thus decrease the performance of our system. We further summarise additional and detailed data-imbalance issues and failure case analysis in Supplementary G.

To further evaluate the segmentation output, we compare the segmentation output of the proposed approach against the ground-truth. Fig. 7 shows that the proposed CANet can effectively predict the correct regions including small tumor cores and complicated edges while the other state of the art methods fail to do so. In Supplementary D, Fig. S4 presents the exemplar segmentation result and the ground-truth annotation in 3D visualisation. From Supplementary D Fig. S4, we observe that our proposed CANet effectively captures 3D forms and shape information in all different circumstances.

Fig. S5 of Supplementary E reports the training curve of CANet and the other State-of-The-Art methods. Our proposed method converges to a lower training loss using less epochs, compared against the other methods. Taking the advantage of the powerful feature interaction graph and the proposed fusion module CGA-CRF, CANet achieves satisfactory outlining of brain glioma. With the training epoch increasing, CANet fine-tunes the segmentation map and successfully detects small

TABLE V
QUANTITATIVE RESULTS OF THE BRATS2018 VALIDATION SET WITH RESPECT TO DICE AND HAUSDORFF95. THE BEST RESULTS OUT OF EACH CATEGORY ARE SHOWN IN BOLD. '-' REPRESENTS THE RESULT OF THE ASSOCIATED METHOD HAS NOT BEEN REPORTED YET. ⋆ AND † REPRESENTS THE FINAL WINNER AND THE RUNNER-UP SOLUTION RESPECTIVELY.

			Dice			Hausdorff95			Model Parameter
Approach	Method		ET	WT	TC	ET	WT	TC	
Ensemble	Isensee et al. [53] †	Mean:	0.796	**0.908**	0.843	3.120	4.790	8.020	1.45E7
		StdDev:	-	-	-	-	-	-	
	McKinley et al. [68]	Mean:	0.793	0.901	0.847	3.603	4.062	**4.988**	-
		StdDev:	-	-	-	-	-	-	
	Zhou et al. [69]	Mean:	0.792	0.907	0.836	2.800	4.480	7.070	-
		StdDev:	-	-	-	-	-	-	
	Cabezas et al. [70]	Mean:	0.740	0.889	0.726	5.304	6.956	11.924	-
		StdDev:	0.277	0.075	0.243	9.964	11.939	13.448	
	Puch et al. [71]	Mean:	0.758	0.895	0.774	4.502	10.656	7.103	1.48E6
		StdDev:	0.264	0.070	0.253	8.227	19.286	7.084	
	Feng et al. [72]	Mean:	0.787	0.906	0.834	3.964	**4.018**	5.340	-
		StdDev:	-	-	-	-	-	-	
	Sun et al. [73]	Mean:	**0.805**	0.904	**0.849**	**2.777**	6.327	6.373	-
		StdDev:	-	-	-	-	-	-	
Single Prediction	Carver et al. [74]	Mean:	0.710	0.880	0.770	4.460	7.090	9.570	2.20E7
		StdDev:	0.290	0.080	0.260	8.320	11.570	14.080	
	Chen et al. [27]	Mean:	0.707	0.845	0.731	10.385	11.822	15.066	1.33E7
		StdDev:	0.264	0.100	0.230	21.205	23.610	20.560	
	Salehi et al. [75]	Mean:	0.704	0.822	0.733	9.668	9.610	13.909	1.20E7
		StdDev:	0.289	0.136	0.242	13.757	13.036	14.965	
	Myronenko [26] ⋆	Mean:	**0.816**	**0.904**	**0.860**	**3.805**	**4.483**	8.278	2.01E7
		StdDev:	-	-	-	-	-	-	
	Weninger et al. [76]	Mean:	0.712	0.889	0.758	8.628	6.970	10.910	-
		StdDev:	-	-	-	-	-	-	
	Gates et al. [77]	Mean:	0.678	0.806	0.685	14.523	14.415	20.017	-
		StdDev:	-	-	-	-	-	-	
	CANet(Ours)	Mean:	0.767	0.898	0.834	3.859	6.685	**7.674**	3.34E7
		StdDev:	0.247	0.082	0.167	11.690	10.135	14.981	

TABLE VI
QUANTITATIVE RESULTS OF THE BRATS2019 VALIDATION SET WITH RESPECT TO DICE AND HAUSDORFF95. THE BEST RESULT IS SHOWN IN BOLD AND THE RUNNER-UP RESULT IS UNDERLINED. ⋆ REPRESENTS THE FINAL WINNER SOLUTION.

	Dice			Hausdorff95		
Method	ET	WT	TC	ET	WT	TC
Jiang et al. [78] ⋆	**0.802**	**0.908**	**0.863**	**3.206**	**4.444**	**5.862**
Zhao et al. [79]	0.702	0.893	0.800	4.766	<u>5.078</u>	<u>6.472</u>
Wang et al. [80]	0.737	<u>0.894</u>	0.807	5.994	5.677	7.357
Li et al. [81]	0.771	0.886	0.813	6.033	6.232	7.409
Myronenko et al. [82]	<u>0.800</u>	<u>0.894</u>	0.834	<u>3.921</u>	5.890	6.562
CANet (Ours)	0.759	0.885	<u>0.851</u>	4.809	7.091	8.409

tumor cores and boundaries. We illustrate the probability map of our proposed CANet and 3D UNet in Fig. 9. From Fig. 9, we witness that our CANet can localize the shape contour of the target tumor to achieve precise segmentation, while the standard 3D UNet may lead to uncertainty, e.g. first row (WT probability map) and last row (TC probability map) in Fig. 9. Also the standard U-Net may misclassify healthy surroundings to be tumor tissues, e.g. second row (WT probability map) and thrid row (ET probability map) in Fig. 9.

We further investigate the segmentation results on the BraTS2017, BraTS2018 and BraTS2019 validation sets, where the quantitative result of each patient case is generated from the online evaluation server. The mean and standard deviation results are reported in Tables IV, V and VI. Box plot in Supplementary F - Fig. S6 shows the distribution of the segmentation result among all the patient cases in the validation set. For the BraTS2017 validation set, our proposed CANet with complete CC+GC and 5-iteration CGA-CRF achieves the State-of-The-Art results of mean Dice scores on ET, TC and mean Hasdorff95 score on ET among the single model segmentation benchmarks. Our CANet has the Dice on ET of 0.728 with standard deviation 0.286, higher than the approach reported in [67]. The Dice on TC by CANet is 0.821, which is higher than the runner-up result reported in [66]. The Hausdorff95 on ET of CANet is 5.496, which is much lower than the runner-up generated in [67]. For the BraTS2018 validation set, our proposed CANet achieves the State-of-The-Art result for Hausdorff95, i.e. 7.674, on the tumor core, while the other results are all runner-ups. Note that the method proposed by Myronenko [26] has the best performance using most of the evaluation metrics. In the Myronenko's method, they set up an additional branch of using autoencoder to regularise the encoder backbone by reconstructing the input 3D MRI image. This autoencoder branch greatly enhances the feature extraction capability of the backbone encoder.

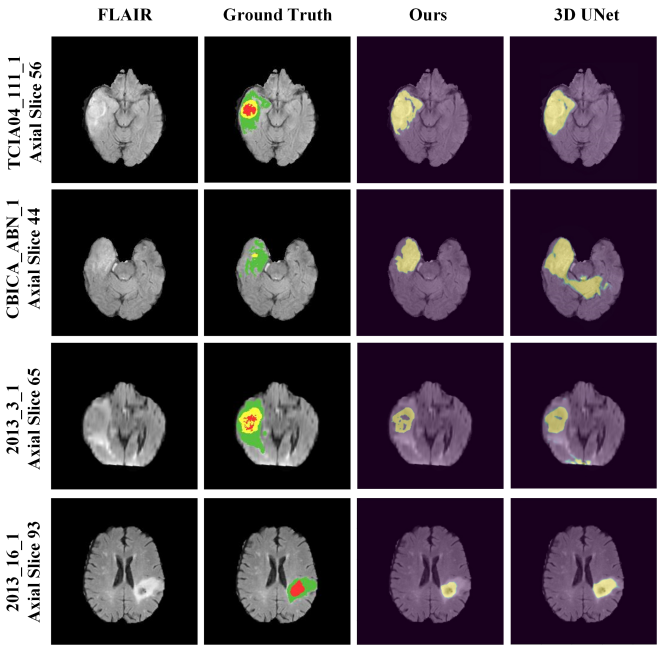


Fig. 9. Examples of segmentation probability maps of our proposed CANet and 3D UNet. Columns from top to bottom represent different patient cases. Rows from left to right indicate the FLAIR data, ground truth annotation, attentive map generated by CANet with 5 iterations in CG-ACRF and attentive map generated by 3D UNet respectively. Best viewed in colors.

In our framework, we regularise the network weights using a L2-regularisation without any additional branch, and the result of our proposed CANet is better than the other single prediction methods. Be reminded that the standard single prediction models generate the segmentation outcomes only using one network, and do not need much computational resources and a complicated voting scheme. For BraTS2019, our proposed CANet with complete CC+GC and 5-iteration CGA-CRF achieves competitive performance against the top performer. CANet's Dice on TC reaches the runner-up and Dice on ET reaches the third place, compared with the other State-of-The-Arts methods [78]–[82]. Note that methods like Jiang et al. [78] used two U-Net in the their single architecture where the first U-Net generates the coarse segmentation result and the second U-Net refines the coarse result to a precise one for the final segmentation output. Thus, this method can be regarded as a variant of model ensembling. Compared with the other State-of-The-Art methods, the result of our proposed CANet is very competitive in terms of accuracy.

Note that even adding additional neural blocks for mean field approximation, the parameters of our system suit the system well. We report the parameter size of our proposed model and other baseline candidates in Tables IV and V (we employ the parameter setup of [83]). Our system maintains the parameter size at a middle level (3.34E7), compared with the other baseline methods such as Islam et al. [60] (1.38E8), Shaikh et al. [62] (2.31E7), Castillo et al. [63] (2.22E7) and Carver et al. [74](2.20E7). We train our system for 200 epoches with a batch size of 2, which takes 27 hours for training. For the evaluation purpose, our system carries out each case within 0.88 seconds.

VI. Conclusion

In summary, we have proposed a novel 3D MRI brain glioma segmentation approach called CANet. Considering different contextual information with standard and graph convolutions, we proposed a novel hybrid context aware feature extractor combined with deep supervised convolution and graph convolution contexts. Different from previous works that used naive feature fusion schemes such as element-wise summation or channel-wise concatenation, we here designed a novel feature fusion model based on conditional random fields, called context guided attentive conditional random field (CGA-CRF), which effectively learns the optimal latent features for downstream segmentation. Furthermore, we formulated the mean-field approximation within CGA-CRF as a convolutional operation, which incorporates the CGA-CRF in a segmentation network to perform end-to-end training. We conducted extensive experiments to evaluate the effectiveness of the proposed feature interaction graph method, CGA-CRF and the complete CANet framework. The results have shown that our proposed CANet achieved the State-of-The-Art results with several evaluation metrics. In the future, we consider combining the proposed network with novel training methods that can better handle the imbalance issue in the datasets.

References

[1] P. M. Black and J. S. Loeffler, *Cancer of the nervous system*. Lippincott Williams & Wilkins, 2005.
[2] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.
[3] A. Mang, S. Bakas, S. Subramanian, C. Davatzikos, and G. Biros, "Integrated biophysical modeling and image analysis: application to neuro-oncology," *Annual review of biomedical engineering*, vol. 22, pp. 309–341, 2020.
[4] S. Bauer, T. Fejes, J. Slotboom, R. Wiest, L.-P. Nolte, and M. Reyes, "Segmentation of brain tumor images based on integrated hierarchical classification and regularization," in *MICCAI BraTS Workshop. Nice: Miccai Society*, 2012.
[5] N. Subbanna and T. Arbel, "Probabilistic gabor and markov random fields segmentation of brain tumours in mri volumes," *Proc MICCAI Brain Tumor Segmentation Challenge (BRATS)*, pp. 28–31, 2012.
[6] H.-C. Shin, "Hybrid clustering and logistic regression for multi-modal brain tumor segmentation," in *Proc. of Workshops and Challanges in Medical Image Computing and Computer-Assisted Intervention (MICCAI'12)*, 2012.
[7] J. Festa, S. Pereira, J. A. Mariz, N. Sousa, and C. A. Silva, "Automatic brain tumor segmentation of multi-sequence mr images using random decision forests," *Proceedings of NCI-MICCAI BRATS*, vol. 1, pp. 23–26, 2013.
[8] S. Reza and K. Iftekharuddin, "Multi-class abnormal brain tissue segmentation using texture," *Multimodal Brain Tumor Segmentation*, vol. 38, 2013.
[9] M. Goetz, C. Weber, J. Bloecher, B. Stieltjes, H.-P. Meinzer, and K. Maier-Hein, "Extremely randomized trees based brain tumor segmentation," *Proceeding of BRATS challenge-MICCAI*, pp. 006–011, 2014.
[10] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.
[11] J. J. Corso, E. Sharon, S. Dube, S. El-Saden, U. Sinha, and A. Yuille, "Efficient multilevel brain tumor segmentation with integrated bayesian model classification," *IEEE transactions on medical imaging*, vol. 27, no. 5, pp. 629–640, 2008.
[12] M. Wels, G. Carneiro, A. Aplas, M. Huber, J. Hornegger, and D. Comaniciu, "A discriminative model-constrained graph cuts approach to fully automated pediatric brain tumor segmentation in 3-d mri," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2008, pp. 67–75.

[13] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in mri images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[14] E. Konukoglu, O. Clatz, P.-Y. Bondiau, H. Delingette, and N. Ayache, "Extrapolating glioma invasion margin in brain magnetic resonance images: Suggesting new irradiation margins," *Medical image analysis*, vol. 14, no. 2, pp. 111–125, 2010.

[15] B. H. Menze, K. Van Leemput, A. Honkela, E. Konukoglu, M.-A. Weber, N. Ayache, and P. Golland, "A generative approach for image-based modeling of tumor growth," in *Biennial International Conference on Information Processing in Medical Imaging*. Springer, 2011, pp. 735–747.

[16] D. Zikic, Y. Ioannou, M. Brown, and A. Criminisi, "Segmentation of brain tumor tissues with convolutional neural networks," *Proceedings MICCAI-BRATS*, pp. 36–39, 2014.

[17] 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.

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

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[20] H. Dong, G. Yang, F. Liu, Y. Mo, and Y. Guo, "Automatic brain tumor detection and segmentation using u-net based fully convolutional networks," in *annual conference on medical image understanding and analysis*. Springer, 2017, pp. 506–517.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[22] M. Lyksborg, O. Puonti, M. Agn, and R. Larsen, "An ensemble of 2d convolutional neural networks for tumor segmentation," in *Scandinavian Conference on Image Analysis*. Springer, 2015, pp. 201–211.

[23] 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.

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

[25] K. Kamnitsas, E. Ferrante, S. Parisot, C. Ledig, A. V. Nori, A. Criminisi, D. Rueckert, and B. Glocker, "Deepmedic for brain tumor segmentation," in *International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries*. Springer, 2016, pp. 138–149.

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

[27] X. Chen, J. Hao Liew, W. Xiong, C.-K. Chui, and S.-H. Ong, "Focus, segment and erase: an efficient network for multi-label brain tumor segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 654–669.

[28] 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.

[29] S. Kumar and M. Hebert, "A hierarchical field framework for unified context-based classification," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, 2005.

[30] T. Toyoda and O. Hasegawa, "Random field model for integration of local information and global information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1483–1489, 2008.

[31] P. Kohli, P. H. Torr *et al.*, "Robust higher order potentials for enforcing label consistency," *International Journal of Computer Vision*, vol. 82, no. 3, pp. 302–324, 2009.

[32] A. Arnab, S. Jayasumana, S. Zheng, and P. H. Torr, "Higher order conditional random fields in deep neural networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 524–540.

[33] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1529–1537.

[34] S. Chen and M. de Bruijne, "An end-to-end approach to semantic segmentation with 3d cnn and posterior-crf in medical images," *arXiv preprint arXiv:1811.03549*, 2018.

[35] M. Shanahan, K. Nikiforou, A. Creswell, C. Kaplanis, D. Barrett, and M. Garnelo, "An explicitly relational neural network architecture," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8593–8603.

[36] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, and D. Rueckert, "Disease prediction using graph convolutional networks: Application to autism spectrum disorder and alzheimer's disease," *Medical image analysis*, vol. 48, pp. 117–130, 2018.

[37] W. Li, Y. Lu, K. Zheng, H. Liao, C. Lin, J. Luo, C.-T. Cheng, J. Xiao, L. Lu, C.-F. Kuo *et al.*, "Structured landmark detection via topology-adapting deep graph learning," *arXiv preprint arXiv:2004.08190*, 2020.

[38] C.-H. Chao, Z. Zhu, D. Guo, K. Yan, T.-Y. Ho, J. Cai, A. P. Harrison, X. Ye, J. Xiao, A. Yuille *et al.*, "Lymph node gross tumor volume detection in oncology imaging via relationship learning using graph neural network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 772–782.

[39] L. Chen, T. Hatsukami, J.-N. Hwang, and C. Yuan, "Automated intracranial artery labeling using a graph neural network and hierarchical refinement," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 76–85.

[40] X. Li, Y. Zhou, N. C. Dvornek, M. Zhang, J. Zhuang, P. Ventola, and J. S. Duncan, "Pooling regularized graph neural network for fmri biomarker analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 625–635.

[41] J. Zhu, L. Fang, and P. Ghamisi, "Deformable convolutional neural networks for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 8, pp. 1254–1258, 2018.

[42] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[43] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl

[44] L. Zhang, X. Li, A. Arnab, K. Yang, Y. Tong, and P. H. S. Torr, "Dual graph convolutional network for semantic segmentation," in *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*. BMVA Press, 2019, p. 254. [Online]. Available: https://bmvc2019.org/wp-content/uploads/papers/0089-paper.pdf

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[46] L. Wang, C.-Y. Lee, Z. Tu, and S. Lazebnik, "Training deeper convolutional networks with deep supervision," *arXiv preprint arXiv:1505.02496*, 2015.

[47] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in neural information processing systems*, 2011, pp. 109–117.

[48] 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, pp. 1993–2024, 2014.

[49] 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.

[50] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.

[51] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[52] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.

[53] 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.

[54] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Atten-

tion u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[55] R. Brügger, C. F. Baumgartner, and E. Konukoglu, "A partially reversible u-net for memory-efficient volumetric image segmentation," *arXiv preprint arXiv:1906.06148*, 2019.

[56] N. Nuechterlein and S. Mehta, "3d-espnet with pyramidal refinement for volumetric brain tumor image segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer International Publishing, 2019.

[57] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, and Y. Fan, "3d brain tumor segmentation through integrating multiple 2d fcnns," in *International MICCAI Brainlesion Workshop*. Springer, 2017, pp. 191–203.

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

[59] A. Jungo, R. McKinley, R. Meier, U. Knecht, L. Vera, J. Pérez-Beteta, D. Molina-García, V. M. Pérez-García, R. Wiest, and M. Reyes, "Towards uncertainty-assisted brain tumor segmentation and survival prediction," in *International MICCAI Brainlesion Workshop*. Springer, 2017, pp. 474–485.

[60] M. Islam and H. Ren, "Multi-modal pixelnet for brain tumor segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2017, pp. 298–308.

[61] 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.

[62] 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.

[63] 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.

[64] 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 *International conference on information processing in medical imaging*. Springer, 2017, pp. 348–360.

[65] 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.

[66] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks," *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 540–549, 2018.

[67] S. Pereira, A. Pinto, J. Amorim, A. Ribeiro, V. Alves, and C. A. Silva, "Adaptive feature recombination and recalibration for semantic segmentation with fully convolutional networks," *IEEE transactions on medical imaging*, 2019.

[68] R. McKinley, R. Meier, and R. Wiest, "Ensembles of densely-connected cnns with label-uncertainty for brain tumor segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 456–465.

[69] C. Zhou, S. Chen, C. Ding, and D. Tao, "Learning contextual and attentive information for brain tumor segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 497–507.

[70] M. Cabezas, S. Valverde, S. González-Villà, A. Clérigues, M. Salem, K. Kushibar, J. Bernal, A. Oliver, and X. Lladó, "Survival prediction using ensemble tumor segmentation and transfer learning," *arXiv preprint arXiv:1810.04274*, 2018.

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

[72] X. Feng, N. Tustison, and C. Meyer, "Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 279–288.

[73] L. Sun, S. Zhang, and L. Luo, "Tumor segmentation and survival prediction in glioma with deep learning," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 83–93.

[74] 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.

[75] S. S. M. 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, 2017.

[76] L. Weninger, O. Rippel, S. Koppers, and D. Merhof, "Segmentation of brain tumors and patient survival prediction: Methods for the brats 2018 challenge," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 3–12.

[77] E. Gates, J. G. Pauloski, D. Schellingerhout, and D. Fuentes, "Glioma segmentation and a simple accurate model for overall survival prediction," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 476–484.

[78] Z. Jiang, C. Ding, M. Liu, and D. Tao, "Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task," in *International MICCAI Brainlesion Workshop*. Springer, 2019, pp. 231–241.

[79] Y.-X. Zhao, Y.-M. Zhang, and C.-L. Liu, "Bag of tricks for 3d mri brain tumor segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2019, pp. 210–220.

[80] F. Wang, R. Jiang, L. Zheng, C. Meng, and B. Biswal, "3d u-net based brain tumor segmentation and survival days prediction," in *International MICCAI Brainlesion Workshop*. Springer, 2019, pp. 131–141.

[81] X. Li, G. Luo, and K. Wang, "Multi-step cascaded networks for brain tumor segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2019, pp. 163–173.

[82] A. Myronenko and A. Hatamizadeh, "Robust semantic segmentation of brain tumor regions from 3d mris," in *International MICCAI Brainlesion Workshop*. Springer, 2019, pp. 82–89.

[83] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, Y. Wang, and Y. Yu, "Exploring task structure for brain tumor segmentation from multi-modality mr images," *IEEE Transactions on Image Processing*, vol. 29, pp. 9032–9043, 2020.

VII. Supplementary B

Fig. S1 summarises the training steps of CANet.

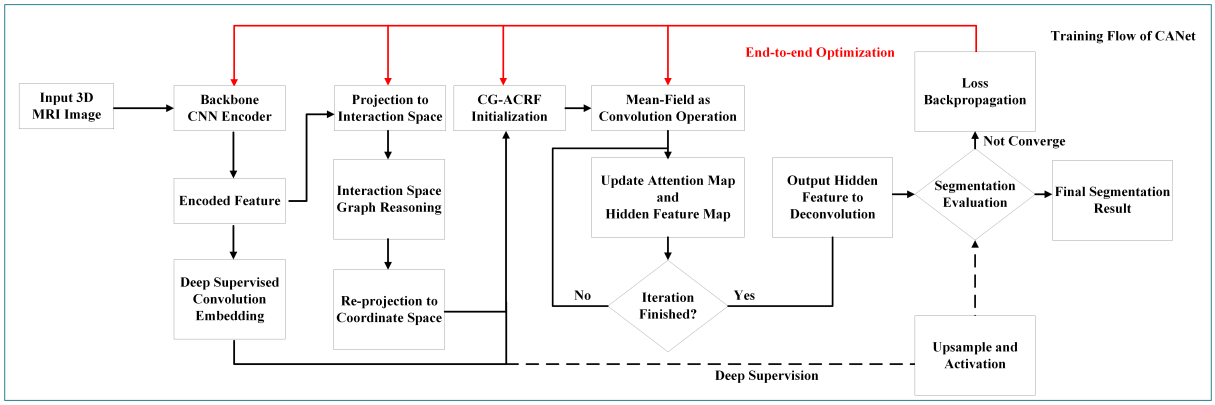


Fig. S1. Training flow of the proposed CANet. Best viewed in colors.

VIII. Supplementary C

Data Augmentation For each sequence in each case, we set all the voxels outside the brain to zero and normalise the intensity of the non-background voxels to be of zero mean and unit variance. During the training, we use randomly cropped images of size 128×128×128. We further set up a common augmentation strategy for each sequence in each case: (i) Randomly rotate an image with the angle between [-20°, +20°]; (ii) Randomly scale an image with a factor of 1.1; (iii) Randomly mirror flip an image across the axial coronal and sagittal planes with the probability of 0.5; (iv) Random intensity shift between [-0.1, +0.1]; (v) Random elastic deformation with $\sigma = 10$.

Implementation Details We implement the proposed CANet and other benchmark experiments using the PyTorch framework and deploy all the experiments on 2 parallel Nvidia Tesla P100 GPUs for 200 epochs with a batch size of 4. We use the Adam optimizer with an initial learning rate $\alpha_0 = 1\mathrm{e}{-4}$. The learning rate is reduced by a factor of 5 after 100, 125 and 150 epochs. We use a $L2$ regulariser with a weight decay of $1\mathrm{e}{-5}$. We store the weights for each epoch and use the weights that lead to the best dice score for inference. The source code will be publicly accessible[4].

IX. Supplementary D

Fig. S2 shows the exemplar segmentation result and the ground truth annotation in 3D visualisation described in Section V-B.

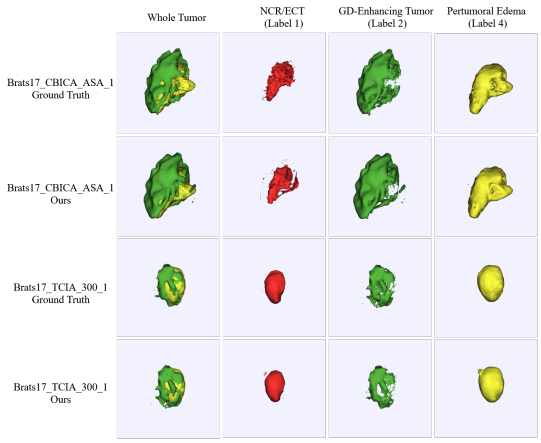


Fig. S2. 3D segmentation results of two volume cases by cross validation on the BraTS2017 training set. The first and third rows indicate the ground truth annotations. The second and fourth rows indicate the segmentation result of our proposed CANet with HCA-FE and 5-iteration CG-ACRF, respectively. Rows from left to right indicate the qualitative comparison of the whole tumor, NCR/ECT, GD-enhancing tumor and Pertumoral Edema respectively. Best viewed in colors.

[4]https://github.com/ZhihuaLiuEd/canetbrats

X. Supplementary E

Fig. S3 reports the training curve of CANet and the other state-of-the-art methods using the BraTS2017 training set, described in Section V-B.

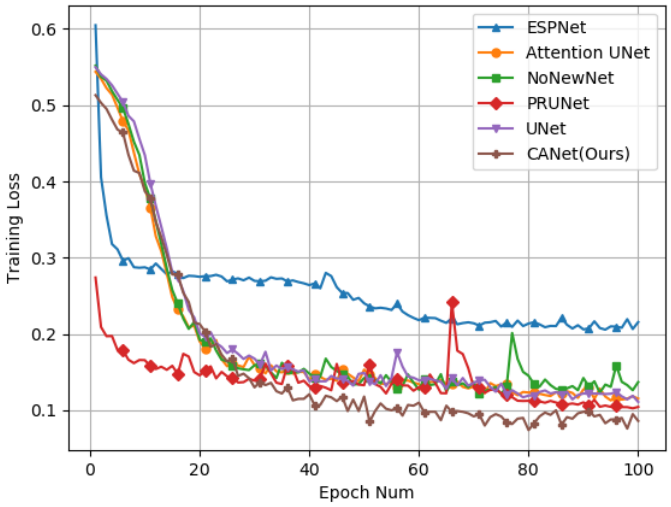


Fig. S3. The learning curve of the state of the art methods and our proposed CANet with HCA-FE and 5-iteration CG-ACRF. Best viewed in color.

XI. Supplementary F

Fig. S4 shows the distribution of the segmentation results among all the patient cases in the BraTS2017 and BraTS2018 validation sets described in Section V-B.

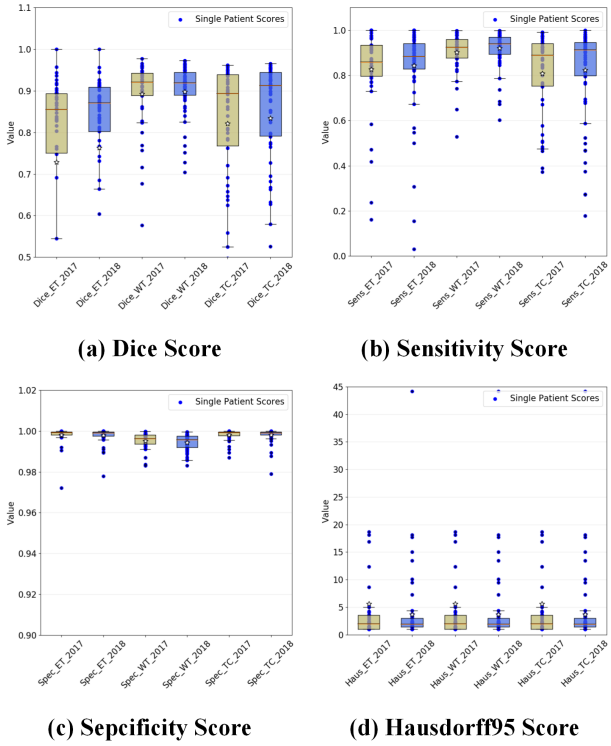


Fig. S4. Boxplot of the segmentation results by CANet with HCA-FE and 5-iteration CG-ACRF, respectively. Dots within yellow boxes are individual segmentation results generated for the BraTS2017 validation set. Dots within blue boxes are individual segmentation results generated for the BraTS2018 validation set. Best viewed in color.

XII. Supplementary G

Fig. S5 shows the statistical information of the BraTS2017 training set. As an example, we here report two failure segmentation cases by our proposed approach, shown in Fig. S6. During the whole training process, CANet focuses on extracting feature maps with different contextual information, e.g. convolutional and graph contexts. However, we have not designed specific strategies for handling the imbalanced issue of the training set. The imbalanced issue is presented in two aspects. Firstly, there exists an unbalanced number of voxels in different tumor regions. As the exemplar case named "Brats17_TCIA_605_1" is shown in Fig. S6, the NCR/ECT region is much smaller than the other two regions, suggesting poor performance of segmenting NCR/ECT. Secondly, there exists an unbalanced number of patient cases from different institutions. This imbalance introduces an annotation bias where some annotations tend to connect all the small regions into a large region while the other annotation tends to label the voxels individually. As the exemplar case named "Brats17_2013_23_1" is shown in Fig. S6, the ground truth annotation tends to be sparse while the segmentation output tends to be connected together. In the future work, we will consider an effective training scheme based on active/transfer learning which can effectively handle the imbalance issue in the dataset. In spite of the imbalance issue, **our segmentation method on the overall cases qualitatively outperforms the other state-of-the-art methods**.

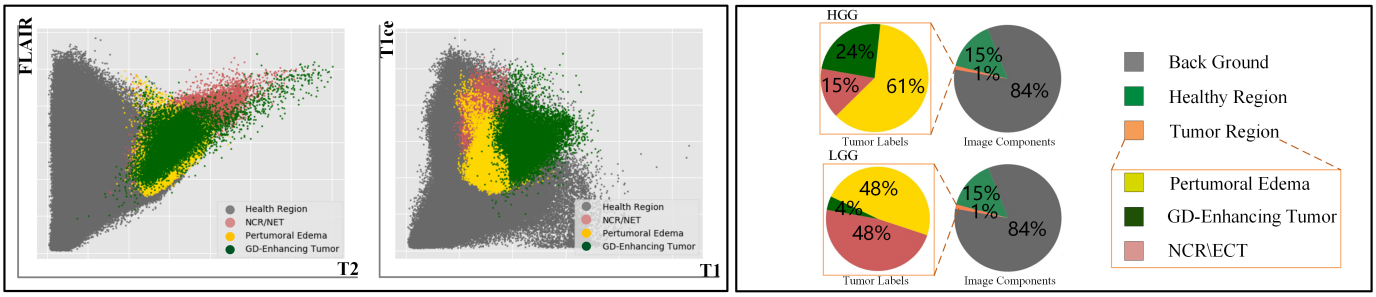


Fig. S5. Statistics of the BraTS2017 training set. The left hand side figure of (a) shows the FLAIR and T2 intensity projection, and the right hand side figure shows the T1ce and T1 intensity projection. (b) is the pie chart of the training data with labels, where the top figure shows the HGG data labels while the bottom figure shows the LGG labels. There are large regions and label imbalance cases here. Best viewed in colors.

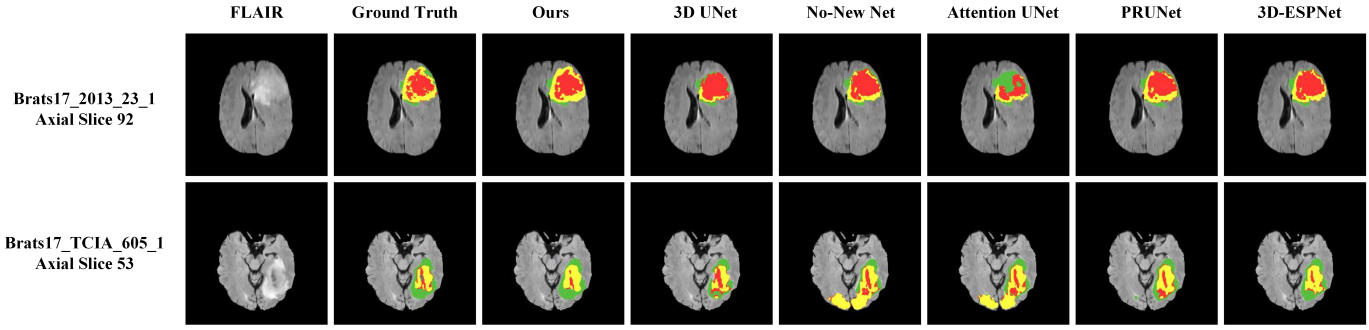


Fig. S6. Qualitative comparisons in the failure cases. Rows from left to right indicate the input data of the FLAIR modality, ground truth annotation, segmentation result from our CANet, segmentation result from the other SOTA methods respectively. Our results look better than the SOTA methods' results. Best viewed in colors.