Inter-slice Context Residual Learning for 3D Medical Image Segmentation

Jianpeng Zhang, Yutong Xie, *Graduate Student Member, IEEE*,
Yan Wang, and Yong Xia, *Member, IEEE*

Abstract—Automated and accurate 3D medical image segmentation plays an essential role in assisting medical professionals to evaluate disease progresses and make fast therapeutic schedules. Although deep convolutional neural networks (DCNNs) have widely applied to this task, the accuracy of these models still need to be further improved mainly due to their limited ability to 3D context perception. In this paper, we propose the 3D context residual network (ConResNet) for the accurate segmentation of 3D medical images. This model consists of an encoder, a segmentation decoder, and a context residual decoder. We design the context residual module and use it to bridge both decoders at each scale. Each context residual module contains both context residual mapping and context attention mapping, the formal aims to explicitly learn the inter-slice context information and the latter uses such context as a kind of attention to boost the segmentation accuracy. We evaluated this model on the MICCAI 2018 Brain Tumor Segmentation (BraTS) dataset and NIH Pancreas Segmentation (Pancreas-CT) dataset. Our results not only demonstrate the effectiveness of the proposed 3D context residual learning scheme but also indicate that the proposed ConResNet is more accurate than six top-ranking methods in brain tumor segmentation and seven top-ranking methods in pancreas segmentation. Code is available at: https://git.io/ConResNet

Index Terms—3D context perceiving, inter-slice context residual, 3D medical image segmentation.

I. INTRODUCTION

MEDICAL images provide visual representations of the anatomy or function of living bodies, which are essential for clinical analysis and medical intervention. Segmentation of 3D medical images aims to predict the semantic category (i.e. a specific organ or lesion) of each voxel, and is a fundamental and critical task in computer aided diagnosis (CAD), since accurate segmentation of organs or tumors is of value not only in facilitating the diagnosis but also in assessing the severity and prognosis of the disease. This segmentation task, however, is extremely challenging due to the low soft tissue contrast and the heterogeneity of organs and tumors in shape, size, and location. Moreover, since the acquisition and

J. Zhang, Y. Xie, Y. Wang, and Y. Xia are with the National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: james.zhang@mail.nwpu.edu.cn; xuyongxie@mail.nwpu.edu.cn; yxia@nwpu.edu.cn).
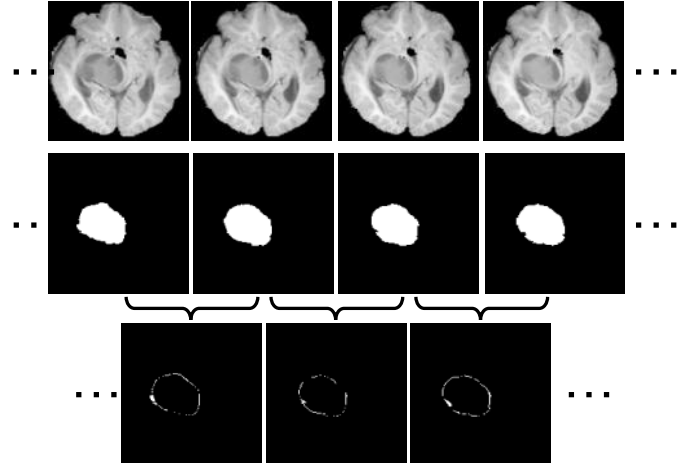

Fig. 1. Illustration of inter-slice context residual in 3D medical images. Top row: adjacent slices of a brain MRI image; middle row: corresponding masks of whole brain tumors; bottom row: inter-slice context residuals between adjacent slices (white: different; black: same).

annotation of medical data are expensive, there is usually a lack of sufficient annotated data to train segmentation models, which renders an even more challenging task for 3D medical image segmentation[1], [2], [3].

Recent years have witnessed the amazing success of deep convolutional neural networks (DCNNs) in image segmentation. Many attempts have been made to strengthen the ability of DCNNs to medical image segmentation. For instance, the encoder-decoder architecture has been improved in various ways to keep the low-level detailed information and obtain sharp object boundaries [4], [5], [6], the spatial pyramid pooling has been used to exploit the multi-scale information [6], the atrous convolution has been incorporated into segmentation models to expand the receptive field efficiently [7], and several attention learning mechanisms have been introduced to segmentation models, enabling them to focus more on certain locations and / or channels [8], [9]. On 3D medical image segmentation tasks, 3D DCNNs have demonstrated striking improvements over their 2D counterparts [10], [11], [12], [13], since they are able to explore the contextual information contained across slices, which contributes significantly to better segmentation performance. We therefore advocate that the accuracy of 3D medical image segmentation can be further improved by capturing and using the inter-slice contextual information more effectively.

However, it is difficult to capture the inter-slice context

information in volumetric images with complex anatomical structures. Take a brain tumor in a magnetic resonance (MR) sequence for example. In Fig. 1, the first row shows four adjacent slices sampled from the sequence in which there is a brain tumor, the second row gives the ground truth tumor region in each slice, and the voxel-wise difference of the tumor region between any two adjacent slices, including extending outward regions or contracting inward regions, is displayed in the third row. We define such difference as the inter-slice context residual, which appears on or near the tumor surface and contains the essential and intriguing morphological information of the tumor, since we can use it, together with the tumor region in any slice, to reconstruct the shape of the 3D tumor. Intuitively, exploring the inter-slice context residual in a segmentation process must be beneficial to improving the accuracy. Unfortunately, since the tumor regions in two adjacent slices have the similar shape and size, the inter-slice context residual is usually tiny, and hence has never been characterized directly.

In this paper, we propose the 3D context residual network (ConResNet) for accurate segmentation of 3D medical images. ConResNet has an encoder-decoder architecture, containing an encoder for feature extraction and two decoders for the generation of segmentation masks and inter-slice context residuals, respectively. The context residual (ConRes) decoder takes the residual feature maps of adjacent slices produced by the segmentation decoder as its input, and also provides feedback to the segmentation decoder as a kind of attention guidance, aiming to boost the ability of the segmentation model to perceive and use the inter-slice context information effectively (see Fig. 2). The design of ConRes decoder is conceptually generic and compatible with any existing 3D DCNN-based medical image segmentation model. We evaluate the proposed ConResNet model on the MICCAI 2018 Brain Tumor Segmentation (BraTS) dataset and NIH pancreas segmentation dataset and achieve the state-of-the-art performance on both segmentation tasks. The main contributions of this paper are summarized as follows:

- We suggest adding the ConRes decoder to an encoder-decoder structure to explicitly boost the model's ability to 3D context perception and thus improve the segmentation accuracy.
- We design the context residual module, which is used between the segmentation decoder and ConRes decoder at each scale, for simultaneous context residual mapping and context attention mapping.
- We propose an accurate 3D medical image segmentation model called ConResNet, which achieves improved performance over state-of-the-art methods on both brain tumor segmentation task and pancreas segmentation task.

II. RELATED WORK

A. Context learning

The context of an object provides the information of its surroundings, and hence is essential for semantic segmentation. Many strategies have been developed for context learning, which can be roughly grouped into three categories. First, to explore the context information at multiple scales, Zhao et al. [14] employed the pyramid spatial pooling strategy at different scales to aggregate multi-scale global information. Chen et al. [6] introduced the atrous spatial pyramid pooling with different dilated rates to parallel branches, which perform multi-scale representation aggregation. Second, to expand the receptive field, Yu et al. [7] replaced the conventional convolution by the atrous convolution with an adjustable dilated rate, which shows superior performance on many computer vision tasks, like segmentation [15], detection[16], and super-resolution[17]. Peng et al. [18] utilized large kernels to capture rich global context information, which is beneficial for dense per-pixel prediction tasks. Third, many attention based methods were proposed to filter out the extraneous information. Hu et al. [19] recalibrated adaptively channel-wise feature responses by explicitly modeling the channel-wise interdependencies of convolutional features. Wang et al. [20] designed an attention module with a bottom-up and top-down feed-forward structure to learn soft attention weights. In our previous work [21], we exploited the intrinsic self-attention ability of deep networks by using the high-level features to generate the attention map for low-level layers.

These strategies have demonstrated their effectiveness in many 2D scenarios, and some of them have been extended to 3D cases. Wang et al. [22] utilized a 3D fully convolutional networks (FCN) with group dilated convolutions to segment the prostate in MR sequences. Schlemper et al. [?] explored a 3D attention gated network which highlights task-related salient regions by embedding the attention mechanism into a U-like segmentation network for 3D organ segmentation. We also attempted to use the 3D atrous spatial pyramid pooling to capture multi-scale information for liver tumor segmentation in CT volumes [23].

Although improving the performance, these 2D and 3D strategies fail to characterize directly the inter-slice context residual information, which is particularly important for 3D medical image segmentation.

B. Residual learning

Residual learning can be traced back to the seminal work by He et al. [24], Since then, it has been successfully applied to many computer vision tasks, including classification, segmentation, detection, and tracking. The idea of residual learning is to learn the residual between the input and output feature maps using the skip connection that directly jumps from input to output. With the skip connection, residual learning eliminates the degradation problem, and hence makes it possible to train very deep networks with up to hundreds, even thousands, of layers. In this work, the proposed ConRes decoder targets at perceiving the feature residual between two adjacent slices, and thus strengthens the model's ability to use the 3D context information for segmentation.

C. Medical image segmentation

Recently, both 2D and 3D DCNNs have become increasingly prevalent tools for medical image segmentation [25], [26], [27], [23], [28], [29]. Ronneberger et al. [4] proposed
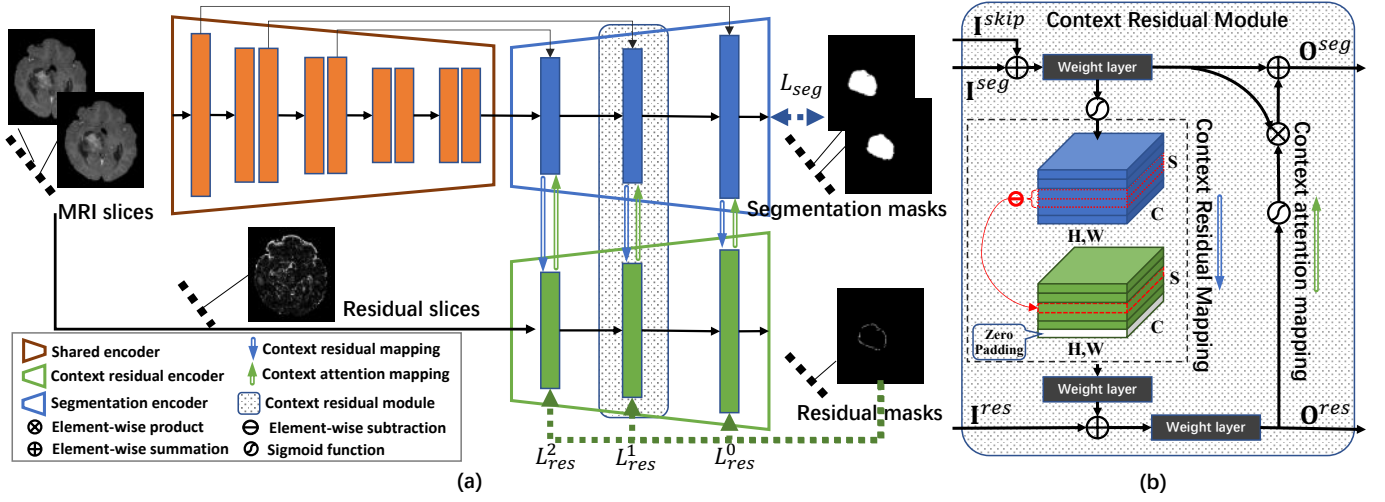
Fig. 2. (a) Diagram of the proposed ConResNet model, which has an encoder-decoder architecture, including a shared encoder (orange), a segmentation decoder (blue), and a ConRes decoder (green). (b) Context residual module.

a U-like architecture (UNet) that consists of an encoder path to capture segmentation-related high-level semantics and a symmetric decoder with skip connections from the encoder to generate segmentation results, and achieved excellent performance on several 2D medical image segmentation tasks. Fu et al. [30] designed a 2D M-Net, which combines a multi-scale U-like network with a side-output layer, and boosted the performance of optic disc and cup segmentation. Chen et al. [31] utilized a 2D deep contour-aware network to harness the multi-level contextual feature representation in an end-to-end way for effective gland segmentation. When applied to 3D medical images, these 2D models perform the segmentation task in a slice-by-slice manner, and hence cannot capture inter-slice context information, leading to limited segmentation accuracy [32].

With 3D convolutions, 3D DCNNs can directly process volumetric data and thus have distinct advantages over 2D DCNNs in 3D medical image segmentation. Dou et al. [11] designed a 3D fully convolutional network to generate high-quality score maps for automated 3D liver segmentation. Li et al. [33] incorporated a multi-scale context module into a 3D segmentation network to use the multi-scale contextual information for inter-vertebral discs segmentation. Meanwhile, Chen et al. [34] integrated the contour length and region constraints into the loss function to enforce the smoothness of segmented regions. Karimi and Salcudean [35] introduced a Hausdorff distance based loss function to minimise the maximum deviation between the prediction and ground truth surfaces of targets. Previously, we developed a light-weight hybrid convolutional network which replaces the 3D convolutions at bottom of network with low-cost 2D convolutions to reduce the model parameters and improve the segmentation performance with only the limited training data [23].

Instead of designing a new segmentation model, we propose to improve the accuracy of existing 3D DCNNs via adding the ConRes decoder to them, which is able to capture and use the inter-slice context residual information.

III. METHOD

Let a 3D medical image be denoted by $\mathbf{X} \in \mathbf{R}^{S \times H \times W}$, where S is the number of slices, and H and W are height and width of each slice, respectively. Its ground-truth segmentation mask is denoted by $\mathbf{Y}_{seg} \in \mathbf{R}^{S \times H \times W}$, in which the semantic label of each voxel is $\mathbf{Y}_{seg}^{s,h,w} = \{0 : background, 1 : object\}$. Its ground-truth context residual mask is denoted by $\mathbf{Y}_{res} \in \mathbf{R}^{S \times H \times W}$, in which each element is calculated as

$$\mathbf{Y}_{res}^{s+1,h,w} = \left| \mathbf{Y}_{seg}^{s+1,h,w} - \mathbf{Y}_{seg}^{s,h,w} \right| \quad (1)$$

where $\mathbf{Y}_{res}^{s+1,h,w} = 0$ means that the voxels at (h, w) on two adjacent (s^{th} and $s+1^{th}$) slices belong to the same semantic categories, i.e., both of them are background or foreground, and $\mathbf{Y}_{res}^{s+1,h,w} = 1$ means that one voxel is background and the other is foreground.

The proposed ConResNet aims to predict the segmentation mask $\mathbf{Y}_{seg}$ and residual mask $\mathbf{Y}_{res}$ simultaneously, formally shown as follows

$$\mathbf{P}_{seg}, \mathbf{P}_{res} = f(\mathbf{X}; \boldsymbol{\theta}) \quad (2)$$

where $\mathbf{P}_{seg} \in \mathbf{R}^{S \times H \times W}$ is the segmentation prediction of $\mathbf{X}$, $\mathbf{P}_{res} \in \mathbf{R}^{S \times H \times W}$ is the residual prediction of $\mathbf{X}$, and $\boldsymbol{\theta}$ represents the parameters of ConResNet. Accordingly, Con-ResNet has a shared encoder for feature extraction and two decoders for the prediction of $\mathbf{Y}_{seg}$ and $\mathbf{Y}_{res}$, respectively. The context residual module is designed to bridge both decoders, which is embedded in each pair of layers of dual decoders. The semantic features from the segmentation decoder are transformed to the context residual features through a new defined operation, called context residual mapping, and input to the ConRes decoder for the refinement. Besides, the ConRes decoder, in turn, provides an attention guidance via the context attention mapping to strengthen the 3D context perceiving ability of the segmentation decoder. The diagram of ConResNet is illustrated in Fig. 2. We now delve into the details of this model.

A. Shared Encoder

In the proposed ConResNet, the shared encoder consists of nine residual blocks, each consisting of two $3 \times 3 \times 3$ convolutional layers and a skip connection from input to output. Due to the limited GPU memory, we have to train the network with very small batch size, like one sample per GPU. Therefore, to speed up the training process, we use the group normalization [36] with a group number of eight, which is insensitive to the scale of batch size, to adjust and scale the activation of each layer. Besides, we use the weight standardization algorithm [37] to accelerate micro-batch training by normalizing the weights of convolutional layers. As shown in Fig. 2, the encoding process can be divided into five stages. In the first stage, the input is processed by a convolutional layer with 32 kernels and a residual block. In each of next three stages, the data is processed by a convolutional layer with doubled kernels and a stride of 2 and two residual blocks. Thus, we gradually down-sample the feature maps to 1/8 of the input size and, simultaneously, increase the number of channels from 32 to 256, leading to the expansion of receptive field and reduction of computation. In the last stage, we employ two residual blocks, which use the atrous convolution with a dilated rate of 2, to further expand the receptive field while keeping the feature resolution for more details of the shape and edge.

B. Dual Decoders

Our ConResNet contains two decoders, i.e., a segmentation decoder and a ConRes decoder. The generation of a segmentation masks by the segmentation decoder or a residual mask by the ConRes decoder consists of three stages. In each stage, we propose a context residual module to bridge the segmentation decoder and ConRes decoder. Specifically, the segmentation decoder first uses the trilinear interpolation to up-sample the previous feature maps, and then fuses them with the low-level features passing from the encoder using element-wise summation. The context residual of fused features are transmitted to the ConRes decoder. The ConRes decoder fuses the context residual features passing from the segmentation decoder and the features passing from the previous layer, and then refine them to predict the residual mask. Moreover, the inter-slice context residual information generated by the ConRes decoder is transmitted back to the segmentation decoder as an attention guidance to boost the ability of the segmentation decoder to perceive 3D context. To match the number of channels in the encoder, we keep halving the channels after each up-sampling operation. As a result, we obtain a segmentation mask and the corresponding inter-slice context residual mask for each input volume.

C. Context Residual Module

Each object in a 3D medical image, such as an organ or a tumor, gradually changes cross-sectional areas on 2D slices. As highlighted by those white pixels in Fig. 3(a), the context residual indicates the discrepancy between the cross-sectional areas on two adjacent slices, which is usually tiny. We suspect
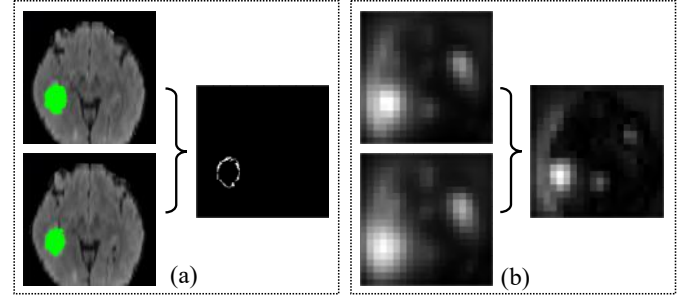

Fig. 3. Illustration of the context residual in images or feature maps. (a) Context residual of brain tumors (green) existing in two adjacent MR slices. (b) Context residual existing in the feature maps (extracted from output of encoder), where the left column gives the channel-wise average of the two adjacent feature maps along the inter-slice dimension. Note that each intensity value in the residual map indicates the level of dissimilarity between two voxels located at the same position in adjacent slices.

that such discrepancy can also be found in the feature maps of adjacent slices produced the segmentation decoder. To verify this, we visualize the channel-wise average of the feature maps of two adjacent slices and their difference in Fig. 3(b)). It shows that those two average feature maps look similar in most regions and the dissimilarity only appear in a small fraction of the slice, which is similar to the pattern of context residual shown in Fig. 3(a). Therefore, we design a context residual module to capture the context residual based on the feature maps produced the segmentation decoder.

Fig. 2(b) shows the diagram of the context residual module. It is composed of two main paths, i.e., a segmentation path (top) and a context residual path (down). In the segmentation path, the input is the element-wise summation of the segmentation features $\mathbf{I}_{seg}$ coming from the previous layer and the low-level features $\mathbf{I}_{skip}$ passing from the same-scale layer in the encoder, and the output is the segmentation feature $\mathbf{O}_{seg}$. In the context residual path, the input is the context residual features $\mathbf{I}_{res}$ coming from the previous layer and the output is the context residual features $\mathbf{O}_{res}$. We design two strategies, i.e., context residual mapping and context attention mapping, to connect two paths for better context perceiving and semantic segmentation.

Context residual mapping: We use a weighted layer to refine the element-wise summation of $\mathbf{I}_{seg}$ and $\mathbf{I}_{skip}$, formally shown as

$$\mathbf{F} = \mathrm{Conv}(\mathbf{I}_{seg} \oplus \mathbf{I}_{skip}), \qquad (3)$$

where $\oplus$ represents the element-wise summation, Conv represents the weighted layer, including convolution, group normalization, and ReLU activation. The obtained feature map $\mathbf{F}$ has a size of $S' \times H' \times W'$, where S', H', W' represents the inter-slice depth, spatial height and width, respectively. The context residual $\mathbf{G}$ is calculated as the position-wise absolute difference between each pair of adjacent features along the inter-slice dimension S, shown as follows

$$\mathbf{G}^{s+1,h,w} = |\sigma(\mathbf{F}^{s+1,h,w}) - \sigma(\mathbf{F}^{s,h,w})|, \qquad (4)$$

where $\sigma(\cdot)$ is the sigmoid function. Note that $\mathbf{G}$ has a size of $(S'-1) \times H' \times W'$. For the convenience of subsequent

processing, we pad the first-slice of $\mathbf{G}$ to make it the same size of $\mathbf{F}$. Then, we combine the generated context residual feature map $\mathbf{G}$ with $\mathbf{I}_{res}$ from the previous context residual layer as the output of context residual path, shown as follows

$$\mathbf{O}_{res} = \text{Conv}(\text{Conv}(\mathbf{G}) \oplus \mathbf{I}_{res}). \qquad (5)$$

Context attention mapping: The context residual mapping generates the inter-slice context residual feature map and feeds it to the residual path for the context residual learning. In the meantime, the output of context residual path can be used as a kind of attention to boost the context perceiving ability of the segmentation path. To this end, we introduce the simple but effective context attention mapping.

We apply the sigmoid function to the context residual feature map $\mathbf{O}_{res}$ and obtain the context attention weights, which is employed as an attention gate to activate the context residual regions and prompt the sensitivity of 3D context variance. Since the context attention highlights only the residual region (see Fig. 2), directly applying the context attention weights to the feature maps may attenuate both the background and foreground. Hence, we define the context attention weighted output of the segmentation path as

$$\mathbf{O}_{seg} = \mathbf{F} \otimes (1 + \sigma(\mathbf{O}_{res})) \qquad (6)$$

where $\otimes$ means the element-wise multiplication.

D. Network Optimization

For the segmentation decoder, the loss function is defined as the following combination of the cross entropy loss and Dice loss

$$L_{seg} = \sum_{s,h,w} l_{bce}(\mathbf{P}_{seg}^{s,h,w}, \mathbf{Y}_{seg}^{s,h,w}) - \frac{2\sum_{s,h,w} \mathbf{P}_{seg}^{s,h,w} \mathbf{Y}_{seg}^{s,h,w}}{\sum_{s,h,w}(\mathbf{P}_{seg}^{s,h,w} + \mathbf{Y}_{seg}^{s,h,w}) + \epsilon} \qquad (7)$$

where l_{bce} is the binary cross entropy loss, $\mathbf{P}_{seg}^{shw}$ and $\mathbf{Y}_{seg}^{shw}$ are the predicted and ground-truth segmentation label at (s, h, w), respectively, and ϵ is a smooth factor in the Dice loss.

Training the ConRes decoder is supervised by the ground-truth context residual mask $\mathbf{Y}^{res}$. The deep supervision technique [11] (see Fig. 2) is adopted to accelerate the convergence. Note that an additional convolutional layer is employed to predict the residual masks for each deep supervision operation. Since the foreground and background voxels are highly imbalanced in each context residual mask, we set a weight for foreground voxels in the binary cross entropy loss function, shown as follows

$$L_{res}^{(i)} = \sum_{s,h,w} w_k l_{bce}(\mathbf{P}_{res}^{s,h,w\,(i)}, \mathbf{Y}_{res}^{s,h,w}) \qquad (8)$$

where $i\ (= 0, 1, 2)$ represents i-th context residual prediction, the class weight $w_k = \log\frac{V}{V_k}$, and V_k is the number of voxels belonging to class k. The overall loss for context residual supervision is defined as

$$L_{res} = L_{res}^0 + \lambda(L_{res}^1 + L_{res}^2) \qquad (9)$$

where the trade-off parameter λ controls the influence of the loss in front layers and is empirically set to 0.5.

Hence, the proposed ConResNet can be jointly optimized by the core segmentation and auxiliary context residual supervisions via minimizing the joint loss $L = L_{seg} + L_{res}$ in an end-to-end manner.

IV. EXPERIMENTS

A. Datasets

Two publicly available datasets were used for this study.
BraTS Dataset: The BraTS dataset [51] was collected and shared by the MICCAI 2018 Brain Tumor Segmentation Challenge [51]. The aim of this challenge is to develop automated segmentation algorithms to delineate intrinsically heterogeneous brain tumors, i.e., (1) enhancing tumor (ET), (2) tumor core (TC) that consists of ET, necrotic and non-enhancing tumor core, and (3) whole tumor (WT) that contains TC and the peritumoral edema. The BraTS dataset has 285 cases for training and 66 cases for online testing. Each case contains four MR sequences, including the T1, T1c, T2, and FLAIR. All sequences were registered to the same anatomical template and interpolated to the same size of $155 \times 240 \times 240$ voxels and the same voxel size of $1.0 \times 1.0 \times 1.0\ mm^3$. The voxel-wise segmentation ground truth of training cases are publicly available, but the ground truth of validation cases are withheld for online evaluation.
Pancreas-CT Dataset: Pancreas-CT[1] is the most authoritative open source dataset for pancreas segmentation, which is collected by the National Institutes of Health Clinical Center [52]. This dataset contains 82 contrasted-enhanced abdominal 3D CT scans from 53 male and 27 female subjects. The size of each scan is $512 \times 512 \times (181 \sim 466)$ voxels and their voxel spacing varies from 0.5mm to 1.0mm. Following [45], [46], [47], [48], [49], [50], we adopted the 4-fold cross-validation and split randomly the dataset into four fixed and roughly equal folds. The average segmentation performance over all 82 cases was reported.

B. Implementation Details

We implemented our ConResNet using PyTorch and performed all experiments on a workstation with two NVIDIA 2080Ti GPUs. During the training, we optimized ConResNet using the Adam algorithm with a batch size of 2 and weight decay of 0.0005. We set the initial learning rate to 0.0001 and decayed it according to a polynomial schedule

$$lr^t = lr^0 \times (1 - t/T)^{0.9} \qquad (10)$$

where t is the index of current iteration, and T is the total number of iterations. To alleviate the overfitting, we employed simple online data augmentation techniques, including randomly scaling and flipping along three dimensions.

On the BraTS dataset, we first normalized the voxel values in each MR sequence to the standard normal distribution, and then concatenated four available modalities into a multimodal volume with 4 channels [42]. In the training stage, we randomly cropped sub-volumes of size $80 \times 160 \times 160$ as training

---

[1]https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT#82af2dca8f2443b1bef1e85ac73acd44

TABLE I
PERFORMANCE OF THE PROPOSED CONRESNET AND SIX STATE-OF-THE-ART SEGMENTATION METHODS ON THE BRATS 2018 ONLINE TEST DATASET.

Methods	Additional data	Dice score (%) / ranking			Hausdorff distance / ranking			Sum-Score
		ET	WT	TC	ET	WT	TC	
CascadeNet [38]	N	79.72 / 7	90.21 / 8	85.83 / 5	3.13 / 6	6.18 / 8	6.37 / 5	39
DMFNet [39]	N	80.12 / 6	90.62 / 5	84.54 / 8	3.06 / 5	4.66 / 6	6.44 / 3	33
OM-Net [40]	N	81.11 / 4	90.78 / 4	85.75 / 6	2.88 / 4	4.88 / 7	6.93 / 8	33
DeepSCAN [41]	N	79.60 / 8	90.30 / 7	84.70 / 7	3.55 / 7	4.17 / 2	4.93 / 1	32
VAE-Seg [42]	N	82.33 / 2	91.00 / 2	86.68 / 1	3.93 / 8	4.52 / 5	6.85 / 7	25
EnsembleNets [43]	N	81.36 / 3	90.95 / 3	86.51 / 2	2.72 / 2	4.17 / 2	6.54 / 6	18
nnUnet [44]	Y	80.87 / 5	91.26 / 1	86.34 / 3	2.41 / 1	4.27 / 4	6.52 / 4	18
ConResNet	N	83.15 / 1	90.38 / 6	85.90 / 4	2.79 / 3	4.15 / 1	5.75 / 2	**17**

TABLE II
COMPARISON OF OUR METHOD AND THE STATE-OF-THE-ART METHODS
ON THE PANCREAS-CT DATASET.

Methods	Mean Dice	Max Dice	Min Dice
HNN [45]	81.27	88.96	50.69
Attention-UNet [8]	83.10	-	-
CNN-RNN [46]	83.70	91.00	59.00
RSTN [47]	84.50	91.02	62.81
ResDSN [48]	84.59	91.45	69.62
Bayesian [49]	85.32	91.47	71.04
Coarse2Fine [50]	85.99	91.20	57.20
Our baseline	85.23	91.90	67.40
ConResNet	**86.06**	**92.00**	**73.40**

samples. We treated this multi-class segmentation problem as three binary segmentation tasks. Accordingly, the last decoding layer of our ConResNet has three output channels, which use the sigmoid activation to produce the segmentation results of ET, WT, and TC, respectively. In this way, we optimized three sub-regions (*i.e.*, ET, WT, and TC) directly.

On the Pancreas-CT dataset, we truncated the HU values of all voxels to the range [-100, +240] to remove irrelevant information [48], and then linearly mapped HU values to [0, 1]. We randomly cropped $64 \times 120 \times 120$ patches as training samples.

Evaluation Metrics: For this study, we adopted the Dice coefficient and Hausdorff distance to measure the performance of all segmentation methods on the BraTS dataset [38], [41], [42], [43], [44], and adopted the mean Dice score, max Dice score, and min Dice score as golden performance indicators on the Pancreas-CT dataset [45], [46], [47], [48], [49], [50].

The Dice coefficient is a statistic used to gauge the overlapping between a segmentation prediction $\mathbf{P}_{seg}$ and the corresponding ground truth $\mathbf{Y}_{seg}$. Specifically, the Dice score is defined as

$$Dice = 2 \times \frac{|\mathbf{P}_{seg} \cap \mathbf{Y}_{seg}|}{|\mathbf{P}_{seg}| + |\mathbf{Y}_{seg}| + \epsilon} \qquad (11)$$

The Hausdorff distance evaluates the quality of segmentation boundaries by computing the maximum distance between the prediction and its ground truth, defined as follows

$$HD = \max\{ \sup_{p \in \partial \mathbf{P}_{seg}} \inf_{y \in \partial \mathbf{Y}_{seg}} \|p - y\|_2 ,$$
$$\sup_{y \in \partial \mathbf{Y}_{seg}} \inf_{p \in \partial \mathbf{P}_{seg}} \|p - y\|_2 \} \qquad (12)$$

where $\partial \mathbf{P}_{seg}$ and $\partial \mathbf{Y}_{seg}$ represents the surface point sets of $\mathbf{P}_{seg}$ and $\mathbf{Y}_{seg}$, respectively. A large Dice coefficient or a small Hausdorff distance indicates a more accurate segmentation result.

C. Results

Comparison on the BraTS dataset: On the BraTS dataset, we compared the proposed ConResNet to six segmentation models top-ranked in the BraTS 2018 Challenge, including CascadeNet [38], DMFNet [39], OM-Net [40], DeepSCAN [41], VAE-Seg [42], EnsembleNets [43], nnUnet [44]. CascadeNet [38] is a triple cascaded network, which hierarchically segments WT, TC and ET. OM-Net [40] separates multi-class segmentation tasks into one deep model and only requires one-pass computation for brain tumor segmentation. DMFNet [39] is built upon a multi-fiber unit, embedding the weighted 3D dilated convolution to gain multi-scale image representation for improved segmentation. DeepSCAN [41] is a densely connected segmentation model that uses dilated convolutions to increase the receptive field. VAE-seg [42] contains a variational auto-encoder to reconstruct the input image itself in order to regularize the segmentation tasks. EnsembleNets [43] is an ensemble of multiple segmentation models, each being specially designed to learn contextual and attentive information. nnUnet [44] incorporates with additional tricks into UNet to boost the segmentation performance.

The Dice scores and Hausdorff distance obtained by using these models to segment three sub-regions, i.e., ET, WT, and TC, were given in TABLE I. Note that the Dice scores and Hausdorff distances of compared methods were reported in the original papers. It shows that our ConResNet achieves the largest Dice score for ET segmentation and smallest Hausdorff distance for WT segmentation, whereas nnUnet achieves the largest Dice score for WT segmentation and smallest Hausdorff distance for ET segmentation. Since no model performs consistently better than others when measured by both metrics on three segmentation tasks, we also give the rank (immediately after the metric score) of each model in terms of each metric on each task, where '1' means performing best and '7' means performing worst. To assess the overall performance of each model, we shows the summation of the ranks over all metrics in the last column, which is treated as the golden performance indicators for brain tumor segmentation. It reveals that our ConResNet outperforms other six state-of-the-art methods on this 3D medical image segmentation
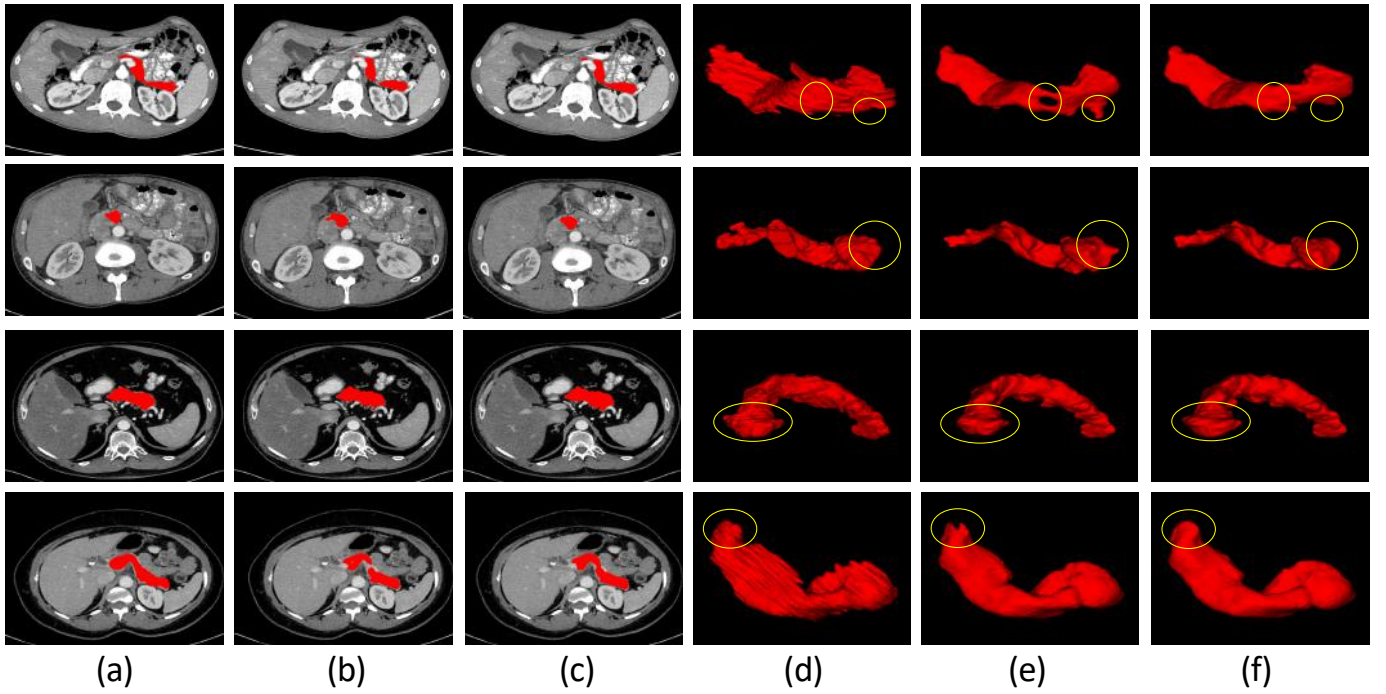
Fig. 4. Comparison of the segmentation results produced by the baseline model and the proposed ConResNet. The left three columns show (a) the ground truth, (b) results of baseline, and (c) results of ConResNet overlaid on CT slices. The right three columns show the 3D visualization of (d) the ground truth, (e) results of baseline, and (f) results of ConResNet. Note that the pancreas is highlighted in red.

problem, as indicated by the smallest Sum-Score. It should be noted that, although achieving an Sum-Score close to ours, nnUnet uses additional data for co-training. Note that all competing methods use the ensemble strategy. nnUnet and VAE-Seg improve the segmentation performance partly due to the ensemble of five and 10 models, respectively. EnsembleNets averages the probabilities predicted by multiple models, including MC-Net, OM-Net and their variants [43]. Thanks to the context residual learning strategy, we achieve competitive results using the ensemble of three ConResNets, which were trained independently.

Comparison on the Pancreas-CT dataset: On the Pancreas-CT dataset, we compared the proposed ConResNet to HNN [45], Attention-UNet [8], CNN-RNN [46], RSTN [47], ResDSN [48], Bayesian [49], and Coarse2Fine [50], which are top-ranked segmentation methods on this task. HNN [45] is a holistically-nested DCNN approach to pancreas localization and segmentation, exploiting multi-view spatial pooling and combining interior and boundary mid-level cues. As an improved version of UNet, Attention-UNet [8] incorporates a novel attention gate module into the model to force it to focus on target structures of varying shapes and sizes. CNN-RNN [46] employs the recurrent neural networks to address the problem of spatial non-smoothness of inter-slice pancreas segmentation along adjacent slices. RSTN [47] repeatedly transforms the segmentation probability map from previous iterations into the spatial prior and use it in the current iteration to relate the coarse and fine stages. ResDSN [48] is composed of a coarse model and a fine model, the former obtains the rough location of the pancreas and the latter

refines the segmentation based on the coarse result. Bayesian [49] leverages the Bayesian model to incorporate the rich shape priors learned from statistical shape models into the deep neural network and thus improves the performance in pancreas segmentation. Coarse2Fine [50] is also a two stage framework for pancreas segmentation, which contains the coarse segmentation for candidate region generation and the fine segmentation of smaller regions-of-interest.

The mean Dice, max Dice, and min Dice of our ConResNet and these methods were listed in TABLE II. Note that the scores of competing methods were adopted from the original papers. It shows ConResNet achieves the highest mean Dice of 86.06%, highest max Dice of 92.00%, and highest min Dice of 73.40%, outperforming the baseline and other competitive methods in terms of all metrics.

Besides, we visualize the segmentation results of the baseline model that does not contain the ConRes decoder and our ConResNet model in Fig. 4. Note that the baseline model has the same encoder and segmentation decoder as ConResNet, including number of channels, network depth, and training strategies. It shows that our ConResNet generates better segmentation results, which are more similar to the ground truth than the results of the baseline model.

Visualization of the learned context residual attention maps: In Fig. 6, we visualized the learned context residual attention maps and compared the feature maps obtained using or without using the context attention mapping. It reveals that the learned context residual attention maps highlight the context residual positions of brain tumor sub-regions. With the context attention mapping, the obtained feature maps have

TABLE III
PERFORMANCE OF OUR CONRESNET WITH / WITHOUT THE CONTEXT RESIDUAL MAPPING AND CONTEXT ATTENTION MAPPING ON THE BRATS VALIDATION DATASET.

Models	Context residual mapping	Context attention mapping	Dice score (%)			Hausdorff distance		
			ET	WT	TC	ET	WT	TC
Baseline	×	×	80.66	89.15	85.89	4.46	5.23	6.30
ConResNet	✓	×	81.15	89.72	87.63	3.39	5.19	5.40
ConResNet	✓	✓	**81.87**	**89.77**	**88.14**	**3.14**	**5.08**	**5.16**

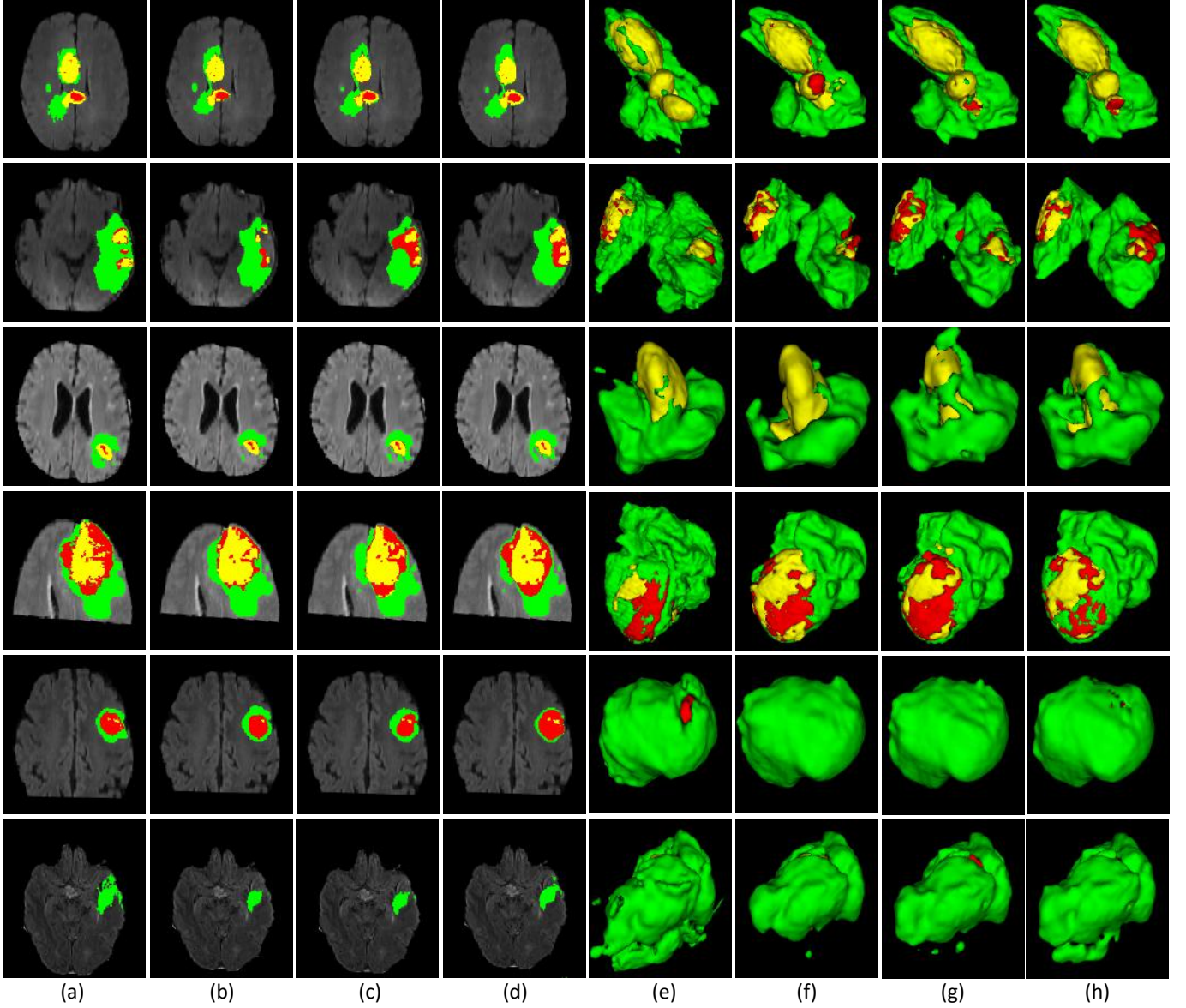
| (a) | (b) | (c) | (d) | (e) | (f) | (g) | (h) |

Fig. 5. Segmentation results obtained by applying our ConResNet with / without the context residual mapping and context attention mapping to six cases on the BraTS validation dataset. From left to right: (a) 2D ground truth overlaid on FLAIR slices, (e) ground truth of 3D tumor, and the 2D and 3D segmentation results of (b,f) the baseline model, (c,g) the baseline with context residual, and (d,h) the baseline with both context residual and context attention (ConResNet). ET: Yellow; TC: Yellow + Red; WT: Yellow + Red + Green.

an improved ability to highlight the boundaries of those sub-regions, which is beneficial for better segmentation.

V. DISCUSSION

A. Ablation Analysis

The major contribution of this work is to add the ConRes decoder and context residual models, which enable both context residual mapping and context attention mapping, to an
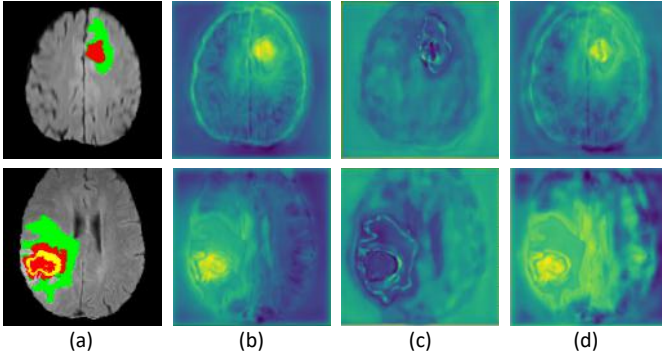
Fig. 6. Visualization of the learned context residual attention maps: (a) input FLAIR MR slices, (b) feature maps obtained without the attention, (c) learned context residual attention maps, and (d) feature maps obtained with the attention. Note that the feature maps shown here are the channel-wise summation of the output of the last context residual module.
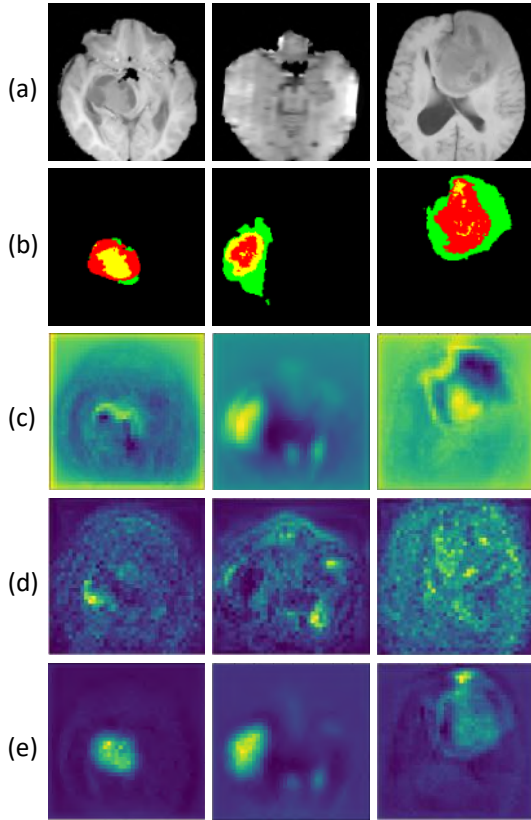

Fig. 7. Visualization of feature maps produced by different models. All the feature maps are extracted from the output of first stage in the decoder. From top to down: (a) T1-weighted brain MR slices as input, (b) ground truth, and feature maps produced by (c) the baseline model, (d) baseline with context residual, and (e) baseline with both context residual and context attention (ConResNet). Note that the feature maps shown here are the channel-wise summation.

encoder-decoder structure for improved segmentation performance. To verify the effectiveness of this design, we performed ablation experiments on the BraTS dataset. For the convenience of quantitative evaluation, we randomly selected 35 cases from training set to form a local validation set. Thus, we have 250 training cases when evaluating on the validation set

(35 cases) and have 285 cases when evaluating on the testing set (66 cases). The performance our ConResNet with / without the context residual mapping and context attention mapping on the local validation set was listed in Table III. Note that we kept other settings such as the network width, depth, and training strategies the same to ensure a fair comparison.

It reveals that using the context residual mapping helps the segmentation model achieve larger Dice scores and smaller Hausdorff distances on ET, WT, and TC, and the performance improvement is substantial in all metrics. Meanwhile, incorporating the context attention mapping into the model with the context residual mapping can further improve the segmentation performance in all metrics. Consequently, comparing to the baseline model that uses neither the context residual mapping nor the context attention mapping, our ConResNet improves the Dice scores by 1.66%, 0.83%, and 2.27% and decreases the Hausdorff distance by 1.76, 0.41, and 1.44 for the segmentation of ET, WT, and TC, respectively.

The segmentation results overlaid on FLAIR slices and the 3D visualization of segmented brain tumors were displayed in Fig. 5. It shows that the results produced by our ConResNet are more similar to the ground truth. Both the quantitative evaluation in Table III and qualitative comparison in Fig. 5 demonstrate the effectiveness of the proposed ConResNet for 3D medical image segmentation.

In Fig. 7, we also visualize the channel-wise summation of the feature maps obtained by the baseline model, the model with only context residual mapping, and our ConResNet that uses both context residual mapping and context attention mapping. It shows that the feature maps produced by our model focus more on the target region than the feature maps produced by other models, which is beneficial for the segmentation.

B. Deep Supervision

The proposed ConResNet also uses deep supervision, which is controlled by the trade-off parameter λ (see Eq.(8)). To validate the contribution made by deep supervision, we trained the model on the BraTS dataset four times when setting λ to 0, 0.1, 0.5, and 1, respectively. Note that $\lambda = 0$ means no deep supervision. In Fig. 8, we observe the lowest training loss and the best performance on the local validation set when $\lambda = 0.5$. Therefore, we empirically set λ to 0.5 in our experiments.

C. Learning Context Residual along Other Views

Our ConResNet is not limited by the inter-slice residual information along the axial view. It is easy to employ the proposed context residual learning to sagittal and coronal views. Taking the sagittal view for example, we compute the position-wise absolute difference between each pair of adjacent features along longitudinal axis, and train our ConResNet to predict the sagittal-view context residual mask. As shown in TABLE IV, we compared the performance of our ConResNet model on the local validation set, considering the residual information along the axial view, sagittal view, and coronal view, respectively. It shows that comparable performance was obtained when considering each of three views. It also reveals that ConResNet with any view is superior to the baseline model in terms of all
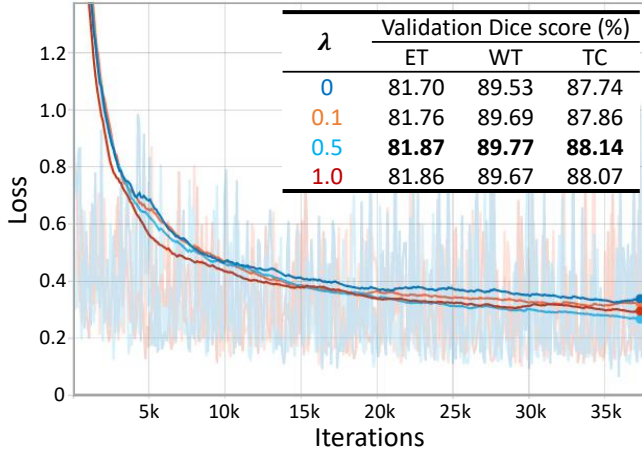
λ	Validation Dice score (%)		
	ET	WT	TC
0	81.70	89.53	87.74
0.1	81.76	89.69	87.86
0.5	**81.87**	**89.77**	**88.14**
1.0	81.86	89.67	88.07

Fig. 8. Plot of training loss of our ConResNet versus iterations when setting λ to different values. The table reports the validation performance under the different setting of λ.

TABLE IV
PERFORMANCE OF CONRESNET WHEN LEARNING CONTEXT RESIDUAL ALONG AXIAL, SAGITTAL, AND CORONAL VIEWS.

Method	View	Dice			Hausdorff distance		
		ET	WT	TC	ET	WT	TC
Baseline	-	80.66	89.15	85.89	4.46	5.23	6.3
ConResNet	Axial	**81.87**	89.77	88.14	3.14	5.08	5.16
ConResNet	Sagittal	81.33	**89.81**	87.78	2.98	**4.51**	5.43
ConResNet	Coronal	81.51	89.70	**88.20**	**2.84**	4.76	**4.63**

evaluation metrics. These results demonstrate the effectiveness of the proposed context residual learning, which is not limited to the axial view.

D. Applying to Multi-class Problems

Let us consider a multi-class segmentation task with C categories, where the ground-truth segmentation mask is denoted by $\mathbf{Y}_{seg} \in \mathbf{R}^{(C+1) \times S \times H \times W}$. For each category $c \in \{1, 2, ..., C\}$, we specifically compute its context residual mask as

$$\mathbf{Y}_{res}^{c,s+1,h,w} = \left| \mathbf{Y}_{seg}^{c,s+1,h,w} - \mathbf{Y}_{seg}^{c,s,h,w} \right| \quad (13)$$

In the multi-class mode, ConResNet predicts the segmentation mask after the $softmax$ activation and predicts the residual mask of each category after the $sigmoid$ activation. Besides, we ignore the residual mask of background and just consider the error back-propagation of C categories during the training process. We compared the multi-class ConResNet to its binary counterpart on the BraTS dataset. The brain tumor segmentation task is formulated either into three binary segmentation problems, including ET vs. others, WT vs. others, and TC vs. others, or into a four-class segmentation problem, where the four classes include: 0 - background, 1 - necrotic and non-enhancing tumor core, 2 - peritumoral edema, and 3 - enhancing tumor. We still evaluated the segmentation performance on three sub-regions as done in the BraTS Challenge. TABLE V shows the performance of the baseline model and

TABLE V
COMPARISON OF CONRESNET AND THE BASELINE MODEL IN BINARY OR MULTI-CLASS MODE. MODE: B - "BINARY"; M - "MULTI-CLASS".

Method	Mode		Dice			Hausdorff distance		
	B	M	ET	WT	TC	ET	WT	TC
Baseline	✓		80.66	89.15	85.89	4.46	5.23	6.30
ConResNet	✓		**81.87**	**89.77**	**88.14**	**3.14**	**5.08**	**5.16**
Baseline		✓	79.85	88.87	85.74	3.48	5.73	5.42
ConResNet		✓	**81.04**	**89.12**	**86.51**	**3.13**	**4.90**	5.37

our ConResNet in either the binary or multi-class mode on the local validation set. It reveals that the binary mode leads to the better performance than the multi-class mode, which may owe to the direct optimization of three sub-regions. However, our ConResNet achieves better performance than the baseline model in both binary and multi-class settings.

E. Comparing to Boundary Loss-based Methods

The methods reported in [35] and [53] use newly designed boundary-wise loss functions to force the model to pay attention to boundary pixels on the object surface. Different intrinsically from them, the proposed ConResNet aims to learn the inter-slice context residual, which contains the essential and intriguing morphological information of the tumor, and use the context residual to boost the segmentation performance via simultaneous context residual mapping and context attention mapping. Although context residual voxels appear on or near to the tumor surface, using the proposed context residual learning is different from defining a boundary loss. In fact, the boundary loss can be incorporated into our model to possibly further improve its performance. In this case, our ConResNet has a compound loss function

$$L_{compound} = \alpha \times L + (1 - \alpha) \times L_{BD} \quad (14)$$

where $\alpha = 1 - t \cdot \frac{1 - 0.01}{T}$ is a weighting factor, and L_{BD} is the boundary loss proposed in [53]. We compared the baseline model and our ConResNet model with and without using the boundary loss on the local validation set. TABLE VI show that using the boundary loss can further improve the performance of the baseline model and our ConResNet. Nevertheless, it also shows that, although the boundary loss is beneficial for this segmentation task, our ConResNet even outperforms the baseline model with the boundary loss. In general, adding a boundary loss or using the context residual learning is able to effectively help the model pay more attention to boundaries so as to improve the segmentation performance, particularly in localizing the surface of each target volume. Interestingly, both approaches address the discrepancies in the boundaries from different perspectives. Jointly using the context residual learning and a boundary loss is able to produce a mutual promotion.

F. Efficiency Analysis

Our ConResNet has dual decoders, which, inevitably, raises the number of parameters and computations. TABLE VII lists

TABLE VI
COMPARISON OF BASELINE AND CONRESNET WITH AND WITHOUT
BOUNDARY LOSS (BD).

Method	Dice score			Hausdorff distance		
	ET	WT	TC	ET	WT	TC
Baseline	80.66	89.15	85.89	4.46	5.23	6.3
Baseline+BD	**81.13**	**89.31**	**86.49**	**4.03**	**5.19**	**6.19**
ConResNet	81.87	89.77	88.14	3.14	5.08	5.16
ConResNet+BD	**81.90**	**89.87**	**88.29**	**2.72**	**4.53**	**4.55**

TABLE VII
NUMBER OF PARAMETERS, MODEL SIZE AND GFLOPS OF DIFFERENT
MODELS.

Models	#Params ($\times 10^6$)	Model size (MB)	GFLOPs
VAE-Seg [42]	23.6	89.7	856.1
Baseline	18.8	75.2	453.9
ConResNet	19.3	77.1	538.5

the number of parameters, model size, Giga Floating-point Operations Per Second (GFLOPs) of VAE-Seg, baseline and our ConResNet. Note that the GFLOPs was calculated when the input is a 3D volume of size $80 \times 160 \times 160$. It shows that our ConResNet suffers from slightly more parameters (an increase of 2.66%), more computations (an increase of 18.64%), and a larger mode size (an increase of 2.53%) than the baseline. However, considering the performance improvement shown in TABLE III and Fig. 5, we believe that such a moderate increase of the complexity is acceptable. Moreover, although it is somewhat slower than the baseline model, our ConResNet can perform brain tumor segmentation almost in real-time, and hence has the potential to be used in clinical practice.

VI. CONCLUSION

In this paper, we propose the ConResNet with explicit 3D context learning to boost the ability of DCNNs to perceive inter-slice context for accurate segmentation of volumetric medical images. We have evaluated this model on the BraTS dataset and Pancreas-CT dataset. Our results indicate that the proposed ConResNet outperforms state-of-the-art method on both brain tumor segmentation and pancreas segmentation tasks. Our ablation study also demonstrates the effectiveness of the proposed context learning, including the context residual mapping and context attention mapping. In our future work, we plan to treat the predicted residual mask as a prior that highlights the error-prone regions and to concatenate it with the input image as the input of another segmentation network for refined result. Besides, we will investigate how to combine the proposed 3D context learning with self-supervised learning, and thus extend this work to semi-supervised segmentation problems.

REFERENCES

[1] 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.
[2] Y. Xie, J. Zhang, Y. Xia, and C. Shen, "A mutual bootstrapping model for automated skin lesion segmentation and classification," *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2482–2493, 2020.
[3] Y. Xie, J. Zhang, H. Lu, C. Shen, and Y. Xia, "SESV: Accurate medical image segmentation by predicting and correcting errors," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2020.
[4] 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 (MICCAI)*, 2015, pp. 234–241.
[5] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2117–2125.
[6] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.
[7] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *International Conference on Learning Representations (ICLR)*, 2016, pp. 1–13.
[8] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, pp. 197–207, 2019.
[9] A. Sinha and J. Dolz, "Multi-scale guided attention for medical image segmentation," *arXiv preprint arXiv:1906.02849*, 2019.
[10] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, "3D deeply supervised network for automatic liver segmentation from CT volumes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2016, pp. 149–157.
[11] Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin, and P.-A. Heng, "3D deeply supervised network for automated segmentation of volumetric medical images," *Medical Image Analysis*, vol. 41, pp. 40–54, 2017.
[12] S. Chen, K. Ma, and Y. Zheng, "Med3D: Transfer learning for 3D medical image analysis," *arXiv preprint arXiv:1904.00625*, 2019.
[13] Z. Zhou, V. Sodha, M. M. R. Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang, "Models genesis: Generic autodidactic models for 3D medical image analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2019, pp. 384–393.
[14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.
[15] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly and semi-supervised semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7268–7277.
[16] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, "Pyramid dilated deeper convlstm for video salient object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 715–731.
[17] Z. Zhang, X. Wang, and C. Jung, "DCSR: Dilated convolutions for single image super-resolution," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1625–1635, 2018.
[18] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4353–4361.
[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.
[20] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3156–3164.
[21] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2092–2103, 2019.
[22] B. Wang, Y. Lei, S. Tian, T. Wang, Y. Liu, P. Patel, A. B. Jani, H. Mao, W. J. Curran, T. Liu *et al.*, "Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic MRI prostate segmentation," *Medical Physics*, vol. 46, no. 4, pp. 1707–1718, 2019.
[23] J. Zhang, Y. Xie, P. Zhang, H. Chen, Y. Xia, and C. Shen, "Light-weight hybrid convolutional network for liver tumor segmentation," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2019, pp. 4271–4277.

[24] 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 (CVPR)*, 2016, pp. 770–778.

[25] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Transactions on Medical Imaging*, vol. 36, no. 4, pp. 994–1004, 2016.

[26] Wang, Yan and Zhou, Yuyin and Shen, Wei and Park, Seyoun and Fishman, Elliot K and Yuille, Alan L, "Abdominal multi-organ segmentation with organ-attention networks and statistical fusion," *Medical image analysis*, vol. 55, pp. 88–102, 2019.

[27] Wang, Yan and Zhou, Yuyin and Tang, Peng and Shen, Wei and Fishman, Elliot K and Yuille, Alan L, "Training multi-organ segmentation networks with sample selection by relaxed upper confident bound," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2018, pp. 434–442.

[28] H. Jia, Y. Xia, Y. Song, D. Zhang, H. Huang, Y. Zhang, and W. Cai, "3d APA-Net: 3d adversarial pyramid anisotropic convolutional network for prostate segmentation in mr images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 2, pp. 447–457, 2020.

[29] Z. Li and Y. Xia, "Deep reinforcement learning for weakly-supervised lymph node segmentation in ct images," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2020.

[30] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *IEEE Transactions on Medical Imaging*, vol. 37, no. 7, pp. 1597–1605, 2018.

[31] H. Chen, X. Qi, L. Yu, and P.-A. Heng, "DCAN: deep contour-aware networks for accurate gland segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2487–2496.

[32] E. Vorontsov, A. Tang, C. Pal, and S. Kadoury, "Liver lesion segmentation informed by joint liver segmentation," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI)*, 2018, pp. 1332–1335.

[33] X. Li, Q. Dou, H. Chen, C.-W. Fu, X. Qi, D. L. Belavy, G. Armbrecht, D. Felsenberg, G. Zheng, and P.-A. Heng, "3D multi-scale FCN with random modality voxel dropout learning for intervertebral disc localization and segmentation from multi-modality MR images," *Medical Image Analysis*, vol. 45, pp. 41–54, 2018.

[34] X. Chen, B. M. Williams, S. R. Vallabhaneni, G. Czanner, R. Williams, and Y. Zheng, "Learning active contour models for medical image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 632–11 640.

[35] D. Karimi and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," *arXiv preprint arXiv:1904.10030*, 2019.

[36] Y. Wu and K. He, "Group normalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[37] S. Qiao, H. Wang, C. Liu, W. Shen, and A. Yuille, "Weight standardization," *arXiv preprint arXiv:1903.10520*, 2019.

[38] G. Wang, W. Li, S. Ourselin, and T. Vercauteren, "Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation," in *International MICCAI Brainlesion Workshop*, 2018, pp. 61–72.

[39] C. Chen, X. Liu, M. Ding, J. Zheng, and J. Li, "3D dilated multi-fiber network for real-time brain tumor segmentation in MRI," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2019.

[40] C. Zhou, C. Ding, X. Wang, Z. Lu, and D. Tao, "One-pass multi-task networks with cross-task guided attention for brain tumor segmentation," *IEEE Transactions on Image Processing*, vol. 29, pp. 4516–4529, 2020.

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

[42] A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop*, 2018, pp. 311–320.

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

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

[45] H. R. Roth, L. Lu, N. Lay, A. P. Harrison, A. Farag, A. Sohn, and R. M. Summers, "Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation," *Medical Image Analysis*, vol. 45, pp. 94–107, 2018.

[46] J. Cai, L. Lu, F. Xing, and L. Yang, "Pancreas segmentation in CT and MRI images via domain specific network designing and recurrent neural contextual learning," *arXiv preprint arXiv:1803.11303*, 2018.

[47] Q. Yu, L. Xie, Y. Wang, Y. Zhou, E. K. Fishman, and A. L. Yuille, "Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8280–8289.

[48] Z. Zhu, Y. Xia, W. Shen, E. K. Fishman, and A. L. Yuille, "A 3D coarse-to-fine framework for automatic pancreas segmentation," *arXiv preprint arXiv:1712.00201*, vol. 2, 2017.

[49] J. Ma, F. Lin, S. Wesarg, and M. Erdt, "A novel bayesian model incorporating deep neural network and statistical shape model for pancreas segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018, pp. 480–487.

[50] N. Zhao, N. Tong, D. Ruan, and K. Sheng, "Fully automated pancreas segmentation with two-stage 3D convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2019, pp. 201—-209.

[51] 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.

[52] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 556–564.

[53] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. Ben Ayed, "Boundary loss for highly unbalanced segmentation," *Medical Image Analysis*, vol. 67, p. 101851, 2021.