2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Integrating Channel Context Attention and Regional Association Attention for Kidney and Tumor Segmentation

Ying Liu, Hui Cui, Tiangang Zhang, Toshiya Nakaguchi, Ping Xuan*

Abstract— **Automatic segmentation of the kidney and tumor from computed tomography (CT) images is an essential step in precision oncology and personalized treatment planning. Due to the irregular shapes and vague boundaries of kidney and tumor, this is a challenging task. Most of existing methods focused on local features without fully considering the associations between regions and contextual relationships between features. We propose a new segmentation method, CR-UNet, to extract, encode and adaptively integrate multiple layers of relevant features. Since the semantic features of different channels contribute differently to the segmentation of kidney and tumor, we introduce semantic attention mechanism of channels. The regional association attention mechanism is established to integrate the semantic and positional connections between different regions. Ablation studies demonstrate the contributions of semantic associations between deep learning channels, and regional relation modelling. Comparison results with state-of-the-art methods over public dataset demonstrated improved tumor and kidney segmentation performance.**

I. Introduction

Accurate segmentation of the kidney and tumor is essential for the diagnosis of cancer and surgery planning [1]. However, manual delineation of kidney and tumor on hundreds of CT slices is time-consuming and prone to inter- and intra-observer variations. Therefore, it is necessary to automatically detect and segment the kidney and tumor from 3D CT volumes to improve segmentation efficiency and accuracy.

Recently, deep learning technology has been widely used in medical image segmentation. Convolutional neural network (CNN) [2] architectures achieved superior performance in computer vision tasks, including U-Net based architecture such as two-dimensional (2D) U-Net [3] and Pyramid Scene Parsing Network (2D PSPNet) [4]. 2D U-Net extracts features from a single CT slice and 2D PSPNet uses different receptive fields to obtain semantic information in the feature map. However, these two 2D models cannot capture the contextual relationship between CT slices. 3D U-Net [5] replaces 2D convolution by 3D convolution operation to preserve spatial information between CT slices. A 3D Res-UNet [6] model was further proposed, which uses residual blocks in 3D U-Net encoder and decoder to improve feature extraction capacity. nnU-Net [7] enables adaptive configuration of model parameters to different datasets, which achieved wide success on 2019 Kidney Tumor Segmentation Challenge (KiTS19) dataset [8]. MSS U-Net [9] introduces a multi-scale supervision scheme in each decoding layer based on nnU-Net. The above-mentioned 3D models can capture the spatial information between 3D CT slices. The relations across the extracted feature maps, however, are not fully explored. Yang *et al*. [10] proposed an FCN_PPM network that combines the basic 3D FCN with a pyramid pooling module to enhance the feature extraction capability. A boundary-aware network (BA-Net) [11] was proposed to improve the segmentation results near object boundaries. However, these methods focus on extracting local regional features which neglect the relationship between multiple objects. Since the left and right kidneys have the same semantic information, there are similar textures and underlying associations between these regions in the entire 3D volume.

In this work, we propose a new Channel-Region (CR) feature learning model based on nnU-Net to learn and integrate multiple levels of features, including channel semantics and the association between regions in the 3D CT volume. The major contributions of our model are summarized below. Firstly, we propose a regional association attention mechanism to measure and model information associations between multiple regions. Secondly, the semantic attention mechanism at channel level is established, to capture informative features for kidney and tumor segmentation.

II. Method

The proposed CR-UNet is given in Figure 1 with Channel-Region feature learning model in Figure 2. Our CR learning model consists of two major components, including channel-level feature extraction, regional association feature learning. Channel-level feature extraction focuses on using the correlation between channels to obtain semantic attention features. The regional association feature learning captures the regional association features of kidney and tumor. Finally, up-sampling is performed to restore the original resolution to achieve end-to-end pixel-level image segmentation.

A. Basic framework architecture

We use nnU-Net as the basic segmentation architecture, consisting of six encoding layers and six decoding layers. The encoder consists of convolutional layers with a kernel size of $3\times3\times3$, instance normalization layers and the activation layers of LeakyRelu and stride convolution for down-sampling. The decoder includes convolutional layers, instance normalization layers, the activation layers of LeakyRelu and transposed convolution operation perform up-sampling.

B. Channel-level feature learning module

The output from nnU-Net encoder is denoted by feature map $F_{enc} \in \mathbb{R}^{C\times H\times W\times D}$ containing deep semantic information. In order to fully explore the semantic connection between

Ying Liu is with the School of Computer Science and Technology, Heilongjiang University, Harbin, China.

Hui Cui is with the Department of Computer Science and Information Technology, La Trobe University, Melbourne, Australia.

Tiangang Zhang is with the School of Mathematical Science, Heilongjiang University, Harbin, China.

Toshiya Nakaguchi is with the Center for Frontier Medical Engineering, Chiba University, Chiba, Japan.

Ping Xuan, is with the School of Computer Science and Technology, Heilongjiang University, Harbin, China (corresponding author; e-mail: xuanping@hlju.edu.cn).

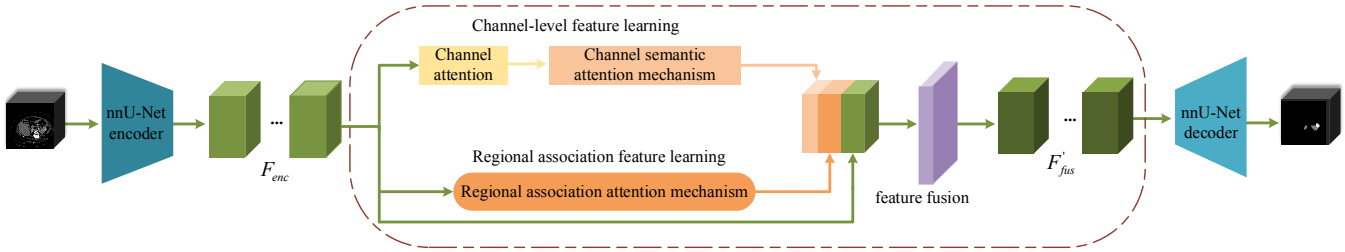

Figure 1. Architecture of the proposed CR-UNet with Channel-Region feature learning model. Details of Channel-Region feature learning model is given in Figure 2.

channels in F_{enc}, we establish the channel-level semantic attention mechanism.

We hypothesize that the features at different positions in F_{enc} are various in different channels, and the semantics of the feature maps at different channels are related to each other. By correlating the feature information between the channels, the features at each position can be enhanced and beneficial to the segmentation of the kidney and the tumor. Channel attention (C-Attention) [12] focuses on the degree of interaction between channels, to model the interdependence between channels. To enhance the feature representation of different channels in F_{enc} for kidney and tumor categories, we use F_{enc} as the input of C-Attention to get $Z_{att} \in \mathbb{R}^{C\times H\times W\times D}$. Z_{att} fusion more information between channel, and also uses correlation to enhance the representation of kidney and tumor features.

Channel semantic attention mechanism. Different channel feature maps have various contributions to kidney and tumor segmentation, so we design a new channel semantic attention. The mechanism learns informative feature representation by giving each channel in feature map z_i a different weight.

We firstly perform 3D convolution operation with $1\times1\times1$ kernel on Z_{att}. The number of convolution kernels is C_s. Let α^k denote the *k-th* convolution kernel, α^k is applied to Z_{att} to obtain $Z_{sem} \in \mathbb{R}^{C_s\times H\times W\times D}$ which is weighted based on the degrees of channel contributions. The process is formally formulated as: $N = H\times W\times D$, and consider N pixels as N nodes. The feature map of *k-th* channel in Z_{sem}, $(Z_{sem})_k$, is calculated as

$$(Z_{sem})_k = \sum_{i=1}^{C} \alpha_i^k * z_i \tag{1}$$

where z_i is the *i-th* channel in Z_{att}, $i \in [1, C]$. α_i^k represents the weight assigned to z_i by *k-th* convolution kernel where $k \in [1, C_s]$. $*$ indicates convolution operation. As shown in Figure 2, each cube represents the features of N nodes from a channel. The operations in the red dashed box in Figure 2(a) represent assigning C weights to the C channel features of the P_i node in Z_{att}.

C. Regional association feature learning module

In the 3D CT volume of a patient, there are correlations between kidney and tumor regions. For instance, there is a certain distance between the left and right kidneys. The kidney tumors exhibit near the kidney. Thus, we propose a regional

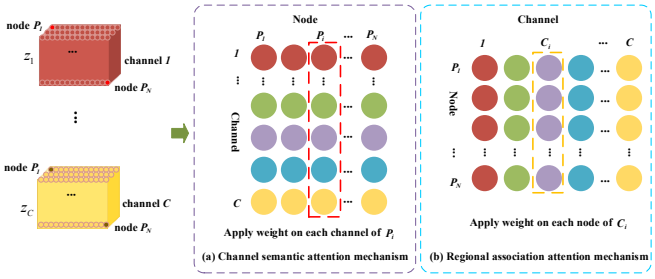

Figure 2. Illustration of Channel-Region feature learning model.

association attention mechanism to model and represent the dependencies of different regions. We firstly perform 3D convolution operation with $1\times1\times1$ kernel on $F_{enc} \in \mathbb{R}^{C\times H\times W\times D}$, the output from nnU-Net encoder, to reduce feature dimension. The output is denoted by $F_{pro} \in \mathbb{R}^{C_r\times H\times W\times D}$ where C_r is the number of reduced channels from C. Secondly, we reshape F_{pro} to $X \in \mathbb{R}^{N\times C_r}$ where $N = H\times W\times D$, each row of X corresponds to a certain image region, and the row vector contains the C_r features of a certain image region. Let X_{i*} represent the feature vector of the *i-th* region. To correlate the features between different regions, we calculate an adjacency matrix B containing the association between regions as

$$B_{ij} = softmax(A_{ij}) = \frac{exp(X_{i*}X_{*j}^T)}{\sum_{i=1}^{N} exp(X_{i*}X_{*j}^T)} \tag{2}$$

where $A = XX^T \in \mathbb{R}^{N\times N}$, *softmax* denotes a softmax layer for row normalization. B records the degrees of mutual influence between N regions, where B_{ij} represents the degree of influence between the features of *i-th* region and *j*-th region. Let $M = BX \in \mathbb{R}^{N\times C_r}$ denote a new matrix that associates features between regions, M_{ic} which represents the features on the *c-th* channel of the *i-th* region is calculated as

$$M_{ic} = \sum_{j=1}^{N} B_{ij} X_{jc} \tag{3}$$

where X_{jc} denotes features on the *c-th* channel in the *j-th* region, B_{ij} denotes the *j-th* region's impact on the *i-th* region. Since regional features representing kidneys or tumors are conducive to the final segmentation, they need to be assigned higher weights. As each row in M corresponds to a regional feature and different regions (different rows) in M have different contributions to kidney and tumor segmentation, we apply 1D convolution to M to adaptively learn the weights of different regions to enhance those regional features which are more important for segmentation. 1D convolution filtered M is obtained as $Z_{reg} \in \mathbb{R}^{N\times C_r}$ where $(Z_{reg})_i$ represents the new

features of the *i-th* region weighted by N regions on different channels

$$(Z_{\text{reg}})_i = \sum_{c=1}^{C_r}\sum_{j=1}^{N}\beta_j^i \bullet \mathrm{M}_{jc} \quad (4)$$

where M_{jc} represents the feature of *j-th* region on the *c-th* channel, β_j^i represents the adaptive weight of the N regions learned by the *i-th* attention mechanism. $\bullet$ represents the dot product operation. The initial value of β_j^i is randomly initialized and automatically learned during training process. Since there are N regions, we set N adaptive attention weights. Finally, we reshape Z_{reg} to get $Z_{reg}^{'} \in \mathbb{R}^{C_r \times H \times W \times D}$ for further processing. Figure 2 shows the difference between regional association attention mechanism and channel semantic attention mechanism. For regional association attention mechanism, the operations in the yellow dashed box in Figure 2(b) indicate that N weights are assigned to the N regional features in the $c_i - th$ channel in M , and the weighted summation leads to the new feature of the $c_i - th$ channel in $(Z_{reg})_i$ .In comparison, the channel semantic attention mechanism applies weights to different channels at the same node, and the regional association attention mechanism applies weights to different points under the same channel.

D. Multi-angle feature fusion and loss function

Given Z_{sem} obtained by the channel semantic attention mechanism, $Z_{reg}^{'}$ obtained by regional association attention mechanism and F_{enc} obtained by nnU-Net encoder, we stack them as $F_{fus} \in \mathbb{R}^{(C_s+C_r+C)\times H\times W\times D}$. F_{fus} is then sent to a 3D convolutional layer with 1×1×1 kernel size to obtain the adaptively fused feature map $F_{fus}^{'} \in \mathbb{R}^{C\times H\times W\times D}$. Finally, the segmentation mask is achieved via nnU-Net decoder.

We use combined loss function of *Dice* [7] and cross entropy (*CE*). Because in 3D CT volume, the tumor region is smaller than the kidney region, we add weight weights ω to kidney and $1-\omega$ to tumor. The hybrid loss function is defined as

$$Loss = (-\log Dice_{kid})^{\xi} \times \omega + (-\log Dice_{tum})^{\psi} \times (1-\omega) + CE \quad (5)$$

where the logarithmic exponential loss operation can increase the punishment intensity of the wrong samples. ξ and ψ are used to control the punishment intensity. *CE* denotes cross entropy loss

$$CE = -\sum_{i\in c}\phi_i y_{true}\log(y_{pred}) \quad (6)$$

where y_{true} represents the ground truth, y_{pred} represents the prediction result, and ϕ_i represents the weight assigned to different categories.

III. Experimental Results

A. Dataset and implementation details

We use 2019 Kidney Tumor Segmentation Challenge (KiTS19) [8] dataset to evaluate the segmentation performance. KiTS19 contains 210 cases of kidney tumors with ground truth. We randomly selected 168 cases from the data and divided them into training set (134 cases) and validation set (34 cases). The remaining 42 cases are used for testing. As metallic artifacts produce abnormal intensity values in CT images, we perform image preprocessing to preserve intensity range between the *0.5th* and *99.5th* percentiles. Then the images are normalized to the range between 0 and 1. The number of slices and voxel spacing in various cases are different. We resample the data to the same voxel spacing of $3.0\times0.78\times0.78\ mm^3$. Patches of size $160\times160\times80$ pixels are extracted from resampled CT volumes for training. Data augmentations include random elastic deformations, random scaling, random rotations, gamma correction augmentation and mirroring. We implemented our method on a single NVIDIA RTX 2080Ti (11 GB RAM) graphic card. We utilize Adam as the network's optimizer function and set the initial learning rate to be 3×10^{-5} , and the batch size is 2. The hyperparameter ω in the loss function is set to 0.4, $\xi = 0.3$, and $\psi = 0.3$. Our model takes 31.93 seconds for segmenting a testing case, while nnU-Net takes 31 seconds. Though our model takes a little more time than nnU-Net, it achieves better segmentation performance.

B. Evaluation measures

Dice is used as evaluation metrics. The *Dice* value is between 0 and 1 where larger value indicates better segmentation results. *Dice* is calculated as follows:

$$Dice = 2\left|D_{pred} \cap D_{true}\right| / \left|D_{pred}\right| + \left|D_{true}\right| \quad (7)$$

where D_{true} represents ground truth, and D_{pred} represents automated segmentation results.

C. Ablation study

To prove the effectiveness of the channel feature learning model (CFL) and regional association feature learning (RAL) modules, we performed ablation studies. In our ablation experiments, we use nnU-Net as the basic framework. The results are given in Table 1. When compared with basic nnU-net, nnU-Net with CFL (nn_CFL) increased Dice of kidney and tumor by 0.4% and 4%, respectively. This is due to the semantic attention mechanism between channels. When RAL was added to nnU-Net (nn_RAL), the Dice of kidney increased by 0.3%, and that of tumor increased to 5.2% dramatically. The results shows that the regional association attention mechanism is beneficial to the segmentation results, especially tumor segmentation. Our CR-UNet outperformed the basic nnU-Net by 0.9% and 6.2% with respect to kidney and tumor dice.

TABLE I. DICE RESULTS OF ABLATION EXPERIMENTS

Method	***Dice* (*kidney*)**	***Dice* (*tumor*)**	***Mean***
nnU-Net	0.958	0.791	0.875
nn_CFL	0.962	0.831	0.897
nn_RAL	0.961	0.843	0.902
CR-UNet	0.967	0.853	0.910

Qualitative evaluation results by nnU-Net, nn_CFL, nn_RAL and CR-UNet of three cases are given in Figure 3. For the case in the first row, the tumor is relatively small and located near the left kidney boundary. We can see that the

shapes of the tumor segmented by nnU-Net, nn_CFL and nn_RAL are much different from the ground truth. The tumor segmented by CR-UNet and ground truth are almost the same. The second row is a case where the tumor is located in the right kidney and occupies a larger region. The right kidney occupies a small area and is not easy to distinguish. nnU-Net segmented only a very small part of the right kidney. With nn_CFL and nn_RAL, the entire kidney was successfully segmented from the tumor and background. The third case is that the tumor is embedded in the left kidney. Obviously, the shape of the tumor segmented by nnU-Net is quite different from the ground truth. Our method achieves the best performance in tumor shape segmentation.

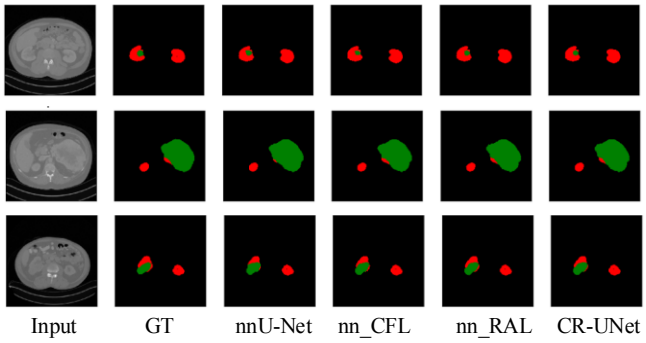

Figure 3. Three cases of segmentation results by nnU-Net, nn_CFL, nn_RAL and CR-UNet. Tumor and kidney segmentation results are shown by green and red.

D. Comparison with other methods

To further evaluate the performance of our method, we compared our network with other methods, including PSPNet [4], 3D U-Net [5], FCN_PPM [10], MSS U-Net [9], 3D Res-UNet [6] and BA-Net [11]. As shown in Table 2, our method achieved the highest Dice value in tumor segmentation, which outperforms the second-best BA-Net by 1.8%, 3D Res-UNet by 2.3%, MSS U-Net by 4.8%, 3D FCN_PPM by 5.1%, 3D U-Net by 10.2%, and 2D PSPNet by 21.5%. BA-Net achieved the second-best performance, mainly because it makes full use of the kidney and tumor's boundary information. The third-best 3D Res-UNet uses the original information, which indirectly proves the necessity of the original information in accurate segmentation. MSS U-Net and 3D FCN_PPM achieve similar performance in tumor segmentation, while MSS U-Net is 3.8% higher than 3D FCN_PPM in kidney segmentation. The performance of 2D PSPNet is the worst. The main reason is that it does not consider the spatial relationship between slices. Our model achieved competitive kidney segmentation results. Besides, our model obtained the highest average Dice values for kidney and tumor, which were 0.6%, 1.2%, 2.3%, 4.3%, 7.1%, and 14% higher than BA-Net, 3D Res-UNet, MSS U-Net, 3D FCN_PPM, 3D U-Net, and 2D PSPNet, respectively.

TABLE II. COMPARISON WITH SIX METHODS ON THE KITS19 CHALLENGE DATASET.

Method	*Dice (kidney)*	*Dice (tumor)*	*Mean*
2D PSPNet [4]	0.902	0.638	0.770
3D U-Net [5]	0.927	0.751	0.839
3D FCN_PPM [10]	0.931	0.802	0.867
MSS U-Net [9]	0.969	0.805	0.887
3D Res-UNet [6]	0.967	0.830	0.898
BA-Net [11]	**0.973**	0.835	0.904
CR-UNet	0.967	**0.853**	**0.910**

IV. CONCLUSION

We propose a new segmentation method, CR-UNet, with channel context attention and regional association attention for kidney and tumor segmentation from 3D CT volumes. Semantic attention mechanism at the channel level distinguishes the contributions between different channels by adaptive weights. The regional association attention mechanism captures the interconnections between different regions. Comparison results and ablation studies on KiTS19 dataset demonstrated the effectiveness of two attentions and improved performance.

ACKNOWLEDGMENT

This work was supported by Degree and Postgraduate Education and Teaching Reform Research Foundation of Heilongjiang University (JGXM_YJS_2019032), Natural Science Foundation of Heilongjiang Province (LH2019A029), Higher Education and Teaching Reform Research Foundation of Heilongjiang University (2019C29), and Natural Science Foundation of China (61972135).

REFERENCES

[1] Taha, Ahmed, et al. "Kid-net: convolution networks for kidney vessels segmentation from ct-volumes." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2018, pp. 463-471.

[2] Krizhevsky A, et al. "ImageNet classification with deep convolutional neural networks." Communications of the ACM 60.6, 2017, pp. 84-90.

[3] Ronneberger, et al. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015, pp. 234-241.

[4] Zhao, Hengshuang, et al. "Pyramid scene parsing network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, pp. 2881-2890.

[5] Çiçek, Özgün, et al. "3D U-Net: learning dense volumetric segmentation from sparse annotation." International conference on medical image computing and computer-assisted intervention. Springer, Cham, 2016, pp. 424-432.

[6] Yu, Lequan, et al. "Volumetric ConvNets with mixed residual connections for automated prostate segmentation from 3D MR images." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 31. No. 1. 2017, pp. 66-72.

[7] Isensee, Fabian, et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation." Nature Methods 18.2, 2021, pp. 203-211.

[8] Heller, Nicholas, et al. "The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes." arXiv preprint arXiv:1904.00445 ,2019, pp. 1-14.

[9] Zhao, Wenshuai, et al. "MSS U-Net: 3D segmentation of kidneys and tumors from CT images with a multi-scale supervised U-Net." Informatics in Medicine Unlocked 19, 2020, pp. 1-11.

[10] Yang, Guanyu, et al. "Automatic segmentation of kidney and renal tumor in ct images based on 3d fully convolutional neural network with pyramid pooling module." 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018, pp. 3790-3795.

[11] Hu, Shishuai, et al. "Boundary-Aware Network for Kidney Tumor Segmentation." International Workshop on Machine Learning in Medical Imaging. Springer, Cham, 2020, pp. 189-198.

[12] Fu, Jun, et al. "Dual attention network for scene segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, pp. 3146-3154.