# Segmentation of kidney lesions with attention model based on Deeplab

AiQing Wen, Xiaochuan Chen, Anni Chen,
Hongyu Shi, and Yuming Hong

Department of Software Engineering,
South China University of Technology, China
https://www.scut.edu.cn

**Abstract.** We participate this challenge by developing a hierarchical framework. We build the model from two fully convolutional networks: (1) a simple Unet model to normalize the input iamges, (2) a segmentaion network which is an attention model based on Deeplab model. Two models are connected in tandem and trained end-to-end. To ensure a better results, we use the preprocess method proposed by nnUnet in our experiments.

**Keywords:** Unet, Deeplab, Attention module

## 1 Introduction

Kidney segmentation from CT volumes is a challenging task due to the low intensity constrast between kidneys and neighboring organs and the voume of kidneys are small compare with other large organ like liver. And kidney tumor segmentation is considered to be a more challending task as the tumor has various size, shape, location and numbers within one patient. And the fuzzy boundary between tumor and noraml tissue has limited the performance of the segmentation methods.

Recent developments of deep neural network have revolutionized the field of computer vision, and have rapidly become a popular mehodology for medical image tasks as well. Current medical image segmentation models are mostly based on fully convolutional neural networks(FCN)[1], often similar to the Unet[2].

To find some other inspiration for model design, we exploit the deeplab model[3–6] which is a architecture that obtain a considerable results on semantic segmentation to construct a model configuration. And a self-attention module is added after the encoder to have the remarkble ability for extracting context information.

## 2 Data Preprocessing

Only the KITS challenge datasets are used for our model training and validation. As for pre-processing,we truncated the voxel values of all CT scans to the range of [-300, 300] HU to eliminate the irrelevant information.

nnUnet [7] is a framework that can automatically adapts itself to any given dataset. We use nnUnet to do the further preprocessing work. And the 2D dataset generator is also used with some modification. The original nnUnet generates single slice image for 2D model, we change it to a 2.5D way instead. We take a stack of adjacent slicces as input and produce the segmentation map corresponding to the center slice. The 2.5D input provides large image content in the axial plane and extra contextual information in the orthogonal direction.

## 3 Method

We build the model from two fully convolutional networks: (1) a simple Unet model to normalize the input iamges, (2) a segmentaion network which is an attention model based on Deeplab model. The overview of the architecture is shown in 1.

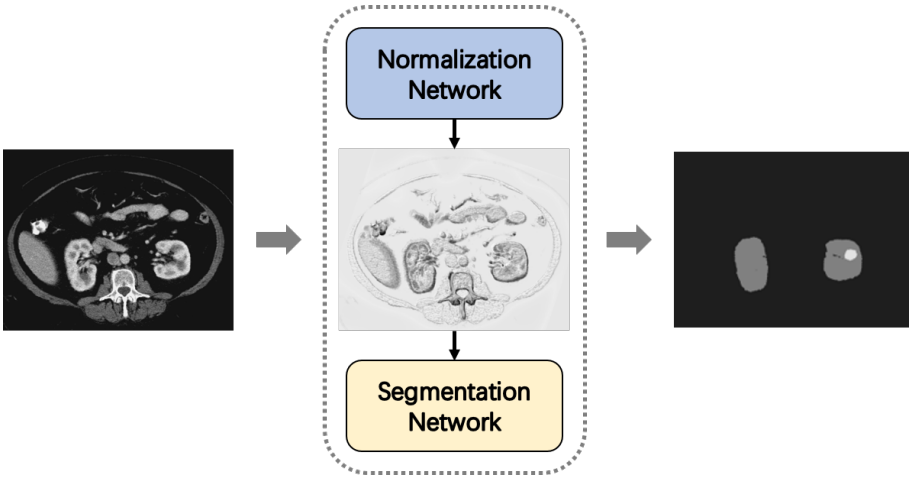


**Fig. 1.** An overview of the architecture.

### 3.1 Normalization network

Research has showed that a low-capacity FCN model can serve as a pre-processor to normalize medical input data [8]. In our experiments, we use a variation of the Unet model as a normalization network (see Fig. 2 for details). The contraction

path is composed of alternating convolution and max-pooling operation, while the extension path is composed of alternating convolution and repeated operation. The extended path restores the lost spatial information in pool operation by connecting the corresponding feature mapping from the contraction path.

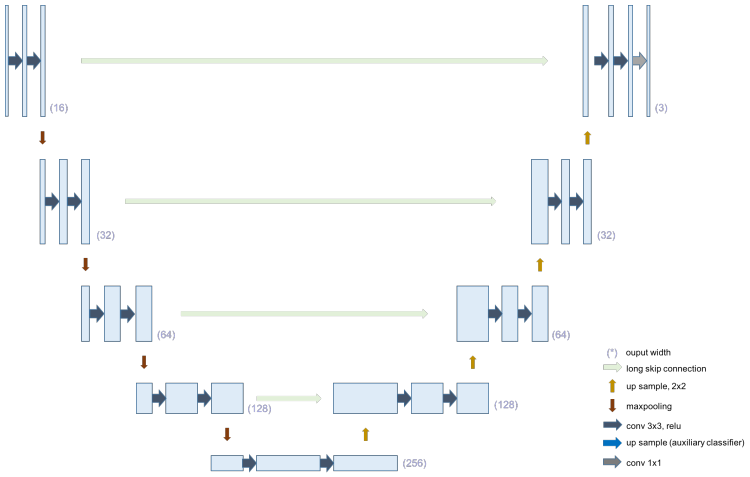


**Fig. 2.** An overview of the Normalization Network.

### 3.2 Segmentation network

We use a attention model based on Deeplab v3+ as our segmentation network. ResNet is the backbone of encoder and a self attention module is added after the ASPP layer to extract context information.

**Atrous convolution:** Atrous convolution is a powerful tool with two attractive advantages: explicitly control the resolution of features computed by deep convolutional neural networks and adjust filter's field-of-view. Atrous convolution allows us to apture multi-scale information, generalizes standard convolution operation.

**Atrous Spatial Pyramid Pooling (or ASPP):** Atrous Spatial Pyramid Pooling module is the combination of atrous spatial pyramid pooling module and image-level features which probes the features with filters at multiple sampling rates and effective field-of-views. ASPP has been used in different network architectures and has shown promising results on several segmentation tasks by exploiting multi-scale information.

**Encoder-decoder:** The encoder-decoder networks have been successfully applied to many computer vision tasks. Typically, the encoder-decoder networks contain (1) an encoder module that gradually reduces the feature maps and captures higher semantic information, and (2) a decoder module that gradually recovers the spatial information.

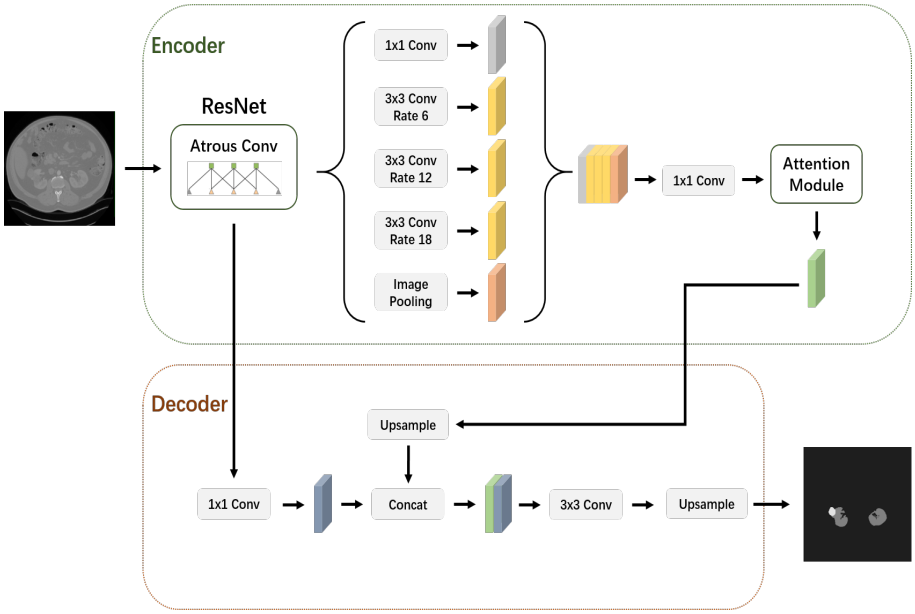


**Fig. 3.** An overview of the Segmentation Network.

**Self-attention module:** Self-attention method is proposed firstly for machine translation. The self-attention module computes the response of a position in the sequence (e.g. a sentence) by focusing on all positions and taking its weighted average in the embedded space. Researches have introduced attention mechanism into visual tasks and achieved state-of-the-art results [9–12]. The self-attention mechanism, see in Fig. 4, can generate stronger pixel-level representation as it enables a single feature from any positions to perceive features from all other positions.

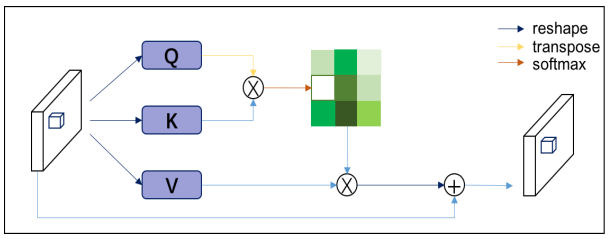


**Fig. 4.** An overview of the self-attention module.

### 3.3 Implementation Details

The model is implemented using the publicly available Pytorch package. The models are trained in a five-fold cross-validation. One epoch is defined as processing 350 batches. Training of each model takes about one day using a single NVIDIA Titan X GPU with 12 GB memory. The model are trained using Adam as optimizer for stochastic gradient descent with an initial learning rate of $3 \times 10^{-4}$ and $l_2$ weight decay of $3 \times 10^{-5}$.

## References


1. E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," IEEE Trans. Pattern Anal. Mach. Intell. (2014).
2. Ronneberger, O., Fischer, P., Brox, T. U-net: Convolutional networks for biomedical image segmentation. in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2015).
3. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. Int. Conf. Learn. Represent. (2016).
4. Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. (2018).
5. Chen, L.-C., Papandreou, G., Schroff, F., Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. (2017).
6. Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2018).
7. Isensee, F., Petersen, J., Kohl, S. A. A., Jäger, P. F., Maier-Hein, K. H. nnU-Net: Breaking the Spell on Successful Medical Image Segmentation. (2019).
8. Drozdzal, M. et al. Learning normalized inputs for iterative estimation in medical image segmentation. Med. Image Anal. 44, 1–13 (2018).
9. Zhao, H. et al. PSANet: Point-wise spatial attention network for scene parsing. in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2018).
10. Huang, Z. et al. CCNet: Criss-Cross Attention for Semantic Segmentation. (2018).
11. Wang, X., Girshick, R., Gupta, A., He, K. Non-local Neural Networks. (2017).
12. Yuan, Y., Wang, J. OCNet: Object Context Network for Scene Parsing. (2018).