A Multi-scale Attention Network for Kidney Tumor Segmentation on CT Scans

Liyan Sun[1], Weihong Zeng[1], Xinghao Ding✉[1], and Yue Huang[1]

Fujian Key Laboratory of Sensing and Computing for Smart City, Xiamen University, Fujian, China dxh@xmu.edu.cn

Abstract. Automatic kidney segmentation is a promising tool for developing advanced surgical planning techniques. However, due to the high morphological heterogeneity within the kidney CT data, the automatic segmentation of kidney and tumor is a difficult problem. Although the state-of-the-art 3D U-Net provides accurate segmentations of medical images, the multi-scale information is underutilized. We propose a *multi-scale attention network* (MSAN) for automatic kidney tumor segmentation. A multi-scale attention layer is developed to combine the local and global contextual information. Furthermore, a ensemble strategy based on voting mechanism boosts the model performance. We achieve the averaged Dice coefficients 94.83% on kidney and 64.89% on tumor in the validation datasets.

Keywords: Kidney Tumor Segmentation · Multi-scale Information · Deep Convolutional Neural Network · Ensemble Learning.

1 Introduction

With more than $400,000$ emerging new cases of kidney cancer each year, the needs for better treatment including surgery is highly desired with the help of automatic semantic segmentation techniques. Since automatic semantic segmentation of kidneys and tumors helps develop surgical planning techniques and predict surgical outcomes, the Kidney Tumor Segmentation (KiTS) 2019 challenge [1] is held up to evaluate the models for automatic kidney and tumor segmentation. Since the kidney and tumor CT images show high morphological diversity, segmenting kidneys and tumor lesions are difficult.

Inspired by the great success of deep convolutional neural network (DCNN) in medical image analysis, the state-of-the-art 3D DCNN model called 3D U-Net [1] achieves accurate segmentation on the volumetric medical images. In this model, the multi-scale information is only exploit by under-sampling or up-sampling the feature maps to exploit global and local contextual information. However, such resize of feature maps causes the spatial information loss, imposing negative effects on the model performance. A more explicit exploitation of multi-scale feature information could lead to better segmentation. Moreover, the multi-scale

[1] https://kits19.grand-challenge.org/home/

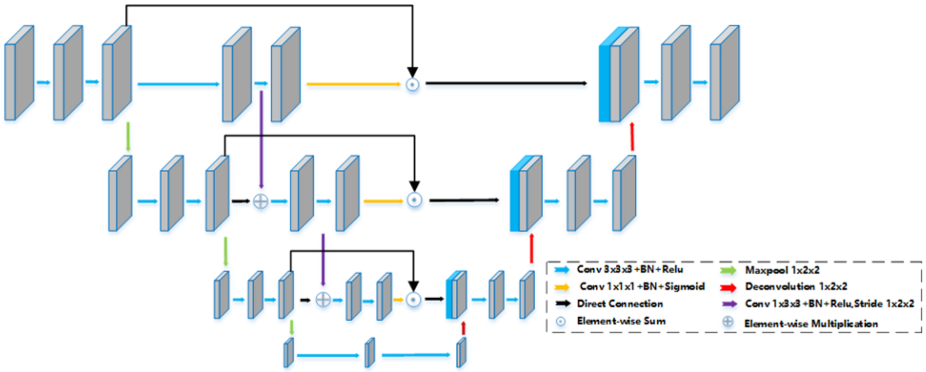


Fig. 1. The MSAN-V1.

information could be incorporated into a attention strategy based on feature recalibration, which elevates the representation ability of the deep networks. Based on the designed multi-scale attention network called MSAN, we further utilize a voting scheme to account for the advantages and disadvantages of each model in the ensembled architecture.

2 Methods

2.1 Data

In the KiTS 2019 challenge, total 300 CT scans of patients who underwent nephrectomy for kidney tumors at the University of Minnesota Medical Center between 2010 to 2018. 210 scans of these patients are picked at random as the training set and released publicly with labeling provided by the challenge organizers. The number of slices in each scan varies from 32 to around 1000. In our experiments, we used the first 180 of the released training sets as our own training set, using ten samples of the 180-190 serial number as our test set. According to statistics, most of the kidneys and tumors have a pixel range between -200 and 500, so we clip them to this interval for all experimental data, which can reduce the impact of other regions on segmentation.

2.2 Model

In figure 1, we show the network flowchart of the multi-scale attention network (MSAN). In this figure, the larger feature maps are added element-wisely with the smaller feature maps in purple arrow, then the information-infused feature maps are input into a Sigmoid layer to generate the weights for feature recalibration. Then a element-wise multiplication is applied for feature enhancement. The enhanced features are then concatenated to the deep layers.

Although the coarse and fine spatial information are fused in the proposed architecture, in the bottom branch of the finest scale, the feature information may not show enough ability to detect the small tumor regions compared with the whole renal regions. To tackle this issue, we propose a MSAN-V2 architecture shown in Figure 2 compared with MSAN-V1 one, adding a 3D Res2Net block [3] parallel to the bottom branch. The architecture of 3D Res2Net block is shown in Figure 3. Different from the explicit multi-scale information fusion architecture design in Figure 1, the exploit of multi-scale information is implicit in the 3D Res2Net block where only the smallest feature maps are input into the block.

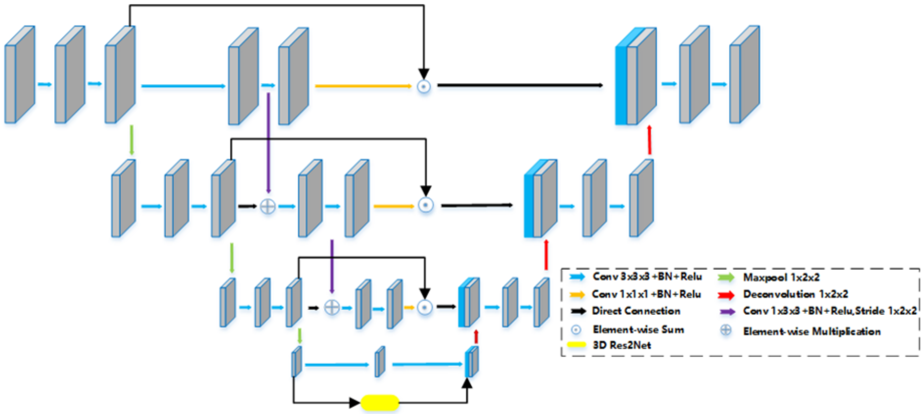


Fig. 2. The MSAN-V2.

Based on the MSAN model and the MSAN-V2 model, in the last two max pooling of the encoding part, we added deep supervision information as part of the model integration. Finally we use a voting strategy among five model of some variants of the MSAN model.

3 Experiments and results

3.1 Implementation

Every 4 images of size 512×512 is input into the trained model with the batchsize set to 2. A dice loss is used as loss function. The Adam optimization method was used to train network for a total of 400,000 training iterations with the initial learning rate of 0.001. We decrease the learning rate by 10 every $100,000$ iterations. We train our model with total $400,000$ iterations. We utilize deep supervision technique [2] in training the models. A simple maximal connected component is applied to further eliminate the worry predictions with isolated regions. Finally, we integrated five models to get better segmentation results,

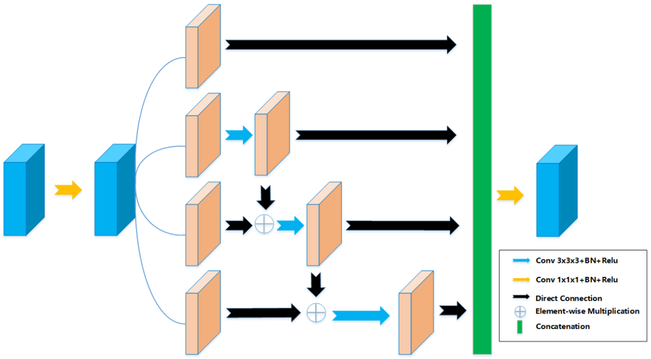


Fig. 3. The 3D Res2Net block.

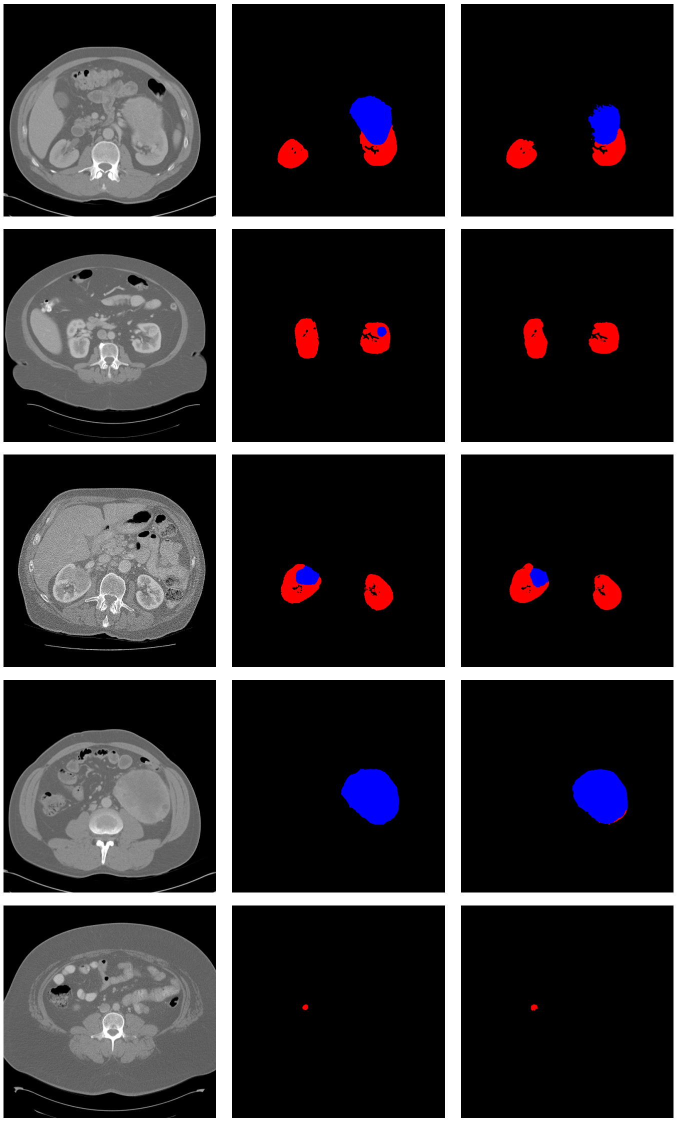
namely 3D-UNET, MSAN-V1, MSAN-V1 with deep supervision, MSAN-V2, MSAN-V2 with deep supervision.

3.2 Results

We report the Dice Coefficient Scores (DC) on the validation datasets in Table 1. We show some example segmentations on figure **??**. As can be seen from the segmentation results, our method achieves better results in the segmentation of the kidney. Whether it is large kidney or small, the segmentation of the edges and details is good. However, due to the multi scale of the tumor, and the contrast with kidney and the edge definition is not clear. The segmentation result is better in the normal size, but there are partial deletions on the abnormal large tumor and sometimes losing smaller tumors.

Model	Kidney	Tumor	Averaged
3D U-Net	89.70	46.90	68.30
MSAN-V1	92.47	56.00	74.23
MSAN-V2	93.85	58.22	76.03
MSAN-V2 with Deep Supervision	93.01	61.41	77.21
Ensembled Model	94.38	64.89	79.63

Table 1. The DS coefficients in percentile % of the compared models on the validation datasets. A simple maximal connected component is applied to all model to further eliminate the worry predictions with isolated regions.



(a) CT image (b) Ground True (c) Predictions

Fig. 4. description of figure

4 Conclusions

In this work, we propose a multi-scale attention network combined with a voting strategy for kidney and tumor segmentation. The rationale behind this architecture is incorporating multi-scale information into feature enhancement idea. The model could be extended to other biomedical image segmentation tasks.

References

1. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: MICCAI. pp. 424–432. Springer (2016)
2. Dou, Q., Chen, H., Jin, Y., Yu, L., Qin, J., Heng, P.A.: 3D deeply supervised network for automatic liver segmentation from CT volumes. In: MICCAI. pp. 149–157. Springer (2016)
3. Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2Net: A new multi-scale backbone architecture. arXiv preprint arXiv:1904.01169 (2019)