Localization Network and End-to-End Cascaded U-Nets for Kidney Tumor Segmentation

Minh H. Vu[1], Guus Grimbergen[2], Attila Simkó[1], Tufve Nyholm[1], and Tommy Löfstedt[1]

[1] Department of Radiation Sciences, Umeå University, Umeå, Sweden
[2] Eindhoven University of Technology, 5612 AZ Eindhoven, the Netherlands

Abstract. Kidney tumor segmentation emerges as a new frontier of computer vision in medical imaging. This is partly due to its challenging manual annotation and great medical impact. Within the scope of the Kidney Tumor Segmentation Challenge 2019, that is aiming at combined kidney and tumor segmentation, this work proposes a novel combination of 3D U-Nets—collectively denoted TuNet—utilizing the resulting kidney masks for the consecutive tumor segmentation. The proposed method achieves a Sørensen-Dice coefficient score of 0.902 for the kidney, and 0.408 for the tumor segmentation, computed from a five-fold cross-validation on the 210 patients available in the data.

1 Introduction

Kidney cancer has an annual worldwide prevalence of over 400 000 new cases, with over 175 000 deaths in 2018 [1]. The most common type of kidney cancer is renal cell carcinoma (RCC) [10]. In Sweden, the indicidence of RCC is 1 125 per 100 000 people, with a 0.75 % risk of developing or dying from the disease [2]. Nephrectomy rates for RCC and metastatic RCC patients increased in Sweden between 2000 and 2008 [13]. Renal tumors can present themselves in a wide variety in terms of size, location and depth [8]. Interest has been developed to study the outcome of partial or radical nephrectomy with respect to tumor morphology [7,9]. For this purpose, but also for reliable disease classification and treatment planning, accurate segmentation of renal tumors in imaging data is of the utmost importance. However, manual annotation is a challenging and time-consuming process in practice, and one that is sensitive to errors and inaccuracies.

In general medical image computer vision, one of the most popular and promising methods in the last years has been Convolutional Neural Networks (CNNs). CNNs learn from data that has been previously segmented, often manually. CNNs have achieved state-of-the-art performance in organ and lesion detection, localization, segmentation, and classification. However, the specific task of kidney and kidney tumor segmentation is challenging due to the aforementioned high degree of variability in tumor location and morphology. As a consequence, the amount of literature that can be found on this topic is limited.

One of the first attempts of using CNNs on this task was by Yang *et al.* (2018) [15] who first used an atlas-based approach to extract two regions of interest from the whole image, each containing one of the kidneys. These patches were used as input to a deep fully connected 3D CNN. After additional post-processing with a conditional random field, an average Sørensen-Dice coefficient (DSC) of 0.931 for the kidney and 0.779 for the tumor were achieved. The Crossbar-Net by Yu *et al.* (2019) [16] used a stack of three rectangular patches in the axial slice, making up what is denoted as a 2.5D data set. These patch shapes allowed for more spatial context to be incorporated, and were used to train two separate networks that combined their respective outcomes in a cascaded approach. This method was designed to segment only the renal tumor, and did so with an average DSC of 0.913.

The present work is developed for the Kidney Tumor Segmentation Challenge 2019 (KiTS19) [3]. Challenges in the field of medical image analysis invite the research community to develop novel approaches to common tasks. Often a data set is made publicly available to the participants, who use it to develop and test their algorithms. The challenge's organizers objectively evaluate and compare the submitted methods. The goal of the KiTS19 challenge is to develop methods that automatically segment kidneys and kidney tumors in computed tomography (CT) images. The KiTS19 challenge is part of the MICCAI 2019 conference. We present here a method that exploits the property of this data set that one structure is a subregion of the other. The encompassing kidney and tumor region is first segmented using a variant of the popular U-Net [11]. The resulting probability map is concatenated with the original image and used as input for a second network, similar in architecture, that predicts the tumor segmentation. Separately, a third network is trained to produce a mask of the whole region, that is fused with the predictions for the whole kidney region and the tumor region. We denote this method TuNet: An End-to-end Hierarchical Tumor Segmentation using Cascaded U-Nets (TuNet).

2   Proposed approach

Our method is a variation of the TuNet that we also used in the Brain Tumors in Multimodal Magnetic Resonance Imaging Segmentation Challenge 2019 (BraTS19). The fundamental idea for our method was based on the work of Wang *et al.* [14] who performed multi-class brain tumor segmentation by a cascade of three networks. Their first network predicted the whole tumor region, which was used as a bounding box to generate the input for the second network. The second network in turn predicted the tumor core area from the input from the first network, and the same method was repeated for a third network that produced a prediction for the enhancing tumor core. The three outputs were finally coalesced into a single multi-label map. For solving other complex, multiple interconnected problems, with a smaller-sized network, cascaded architectures have been proposed before [6]. The combination of two 2.5D networks have sur-
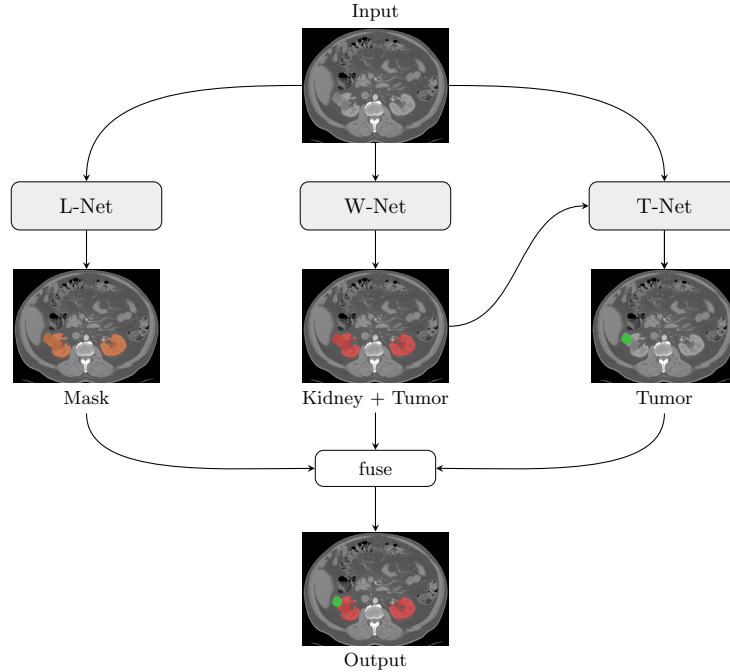
Fig. 1: The proposed Tu-Net model. The input CT image is used in a combination of U-Nets. First independently in the Localization Network (L-Net), and then in the network for whole kidney segmentation (W-Net) and finally using the kidney output probability map together with the original image in the network for the tumor segmentation (T-Net).

passed the accuracy of current state-of-the-art methods in handling automatic detection from 4D microscopic data [6].

For this challenge we employed a similar hierarchical architecture, since the tumor structure is always a sub-volume of the kidney + tumor structure. We first trained a network called W-Net on the whole region, and concatenated the resulting segmentation with the original image. This output was then used as the input for a second network, called T-Net, that produced a mask for the tumor only. Further, we extended the method of Wang *et al.* [14] by separately training a localization network, denoted L-Net, that produced a crude mask of the whole region. This mask was then finally fused by multiplying with the W-Net and T-Net outputs, in order to eliminate false positives and isolated structures that lie outside the main region of interest. A summary of the proposed method is illustrated in Figure 1.
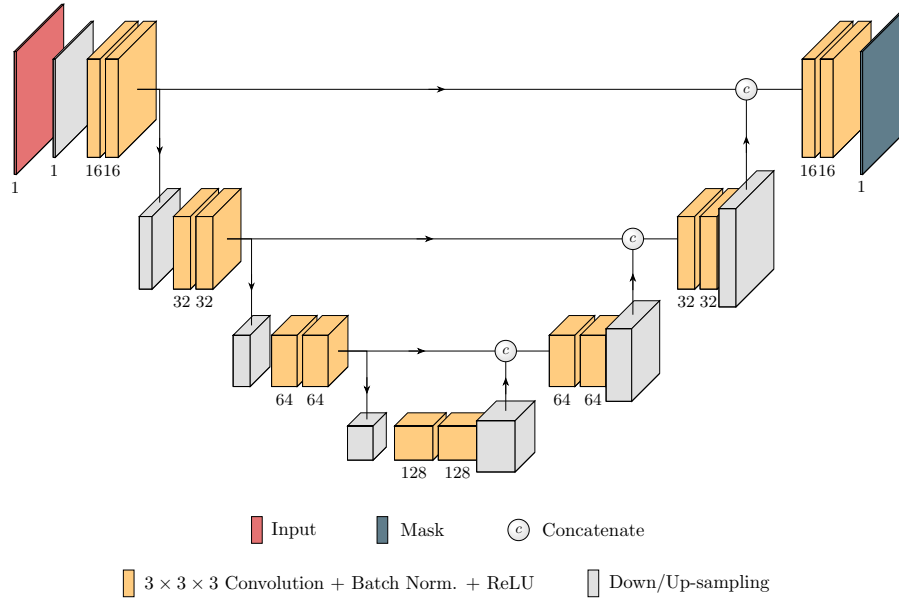
Fig. 2: The Localization Network (L-net) architecture. The figure is simplified to the single-slice case, however the convolutional layers were 3D.

2.1   Localization Network

The localization network, L-Net, was an end-to-end 3D variant of the U-Net, an encoder-decoder architecture that has been proven powerful in semantic medical image segmentation [11]. In order to fit the whole 3D image into the GPU memory at once, the volume was downsampled to the size of $128 \times 128 \times 64$. It was then fed into a traditional 3D U-Net to estimate roughly the region of interest of the whole (kidney + tumor) region. The L-Net is illustrated in Figure 2. For simplicity of illustration, it is shown with a 2D input and output.

2.2   Cascaded Architecture

The model architectures used for the whole region and tumor segmentation had a similar U-Net architecture to the localization network. An additional feature of both the W-Net and the T-Net was that in each convolution block in the encoding path, the original input image was downsampled and concatenated with the feature map of the respective layer. The rate of downsampling was increased according to the dimensions of that layer.

As mentioned before, we employed a cascaded approach. The input of the T-Net consisted of two channels: the input image, and the probability map of the previous output. An illustration of the T-Net architecture is given in Figure 3. The W-Net was analogous, but lacked the probability map as secondary input.

1 1   16 16

↓ 2

32 32

↓ 4

64 64

↓ 8

128   128

16 16   1

32 32

64 64

- Input
- Output
- Down-sampled input
- Probability map of previous network
- $3 \times 3 \times 3$ Convolution + Batch Norm. + ReLU
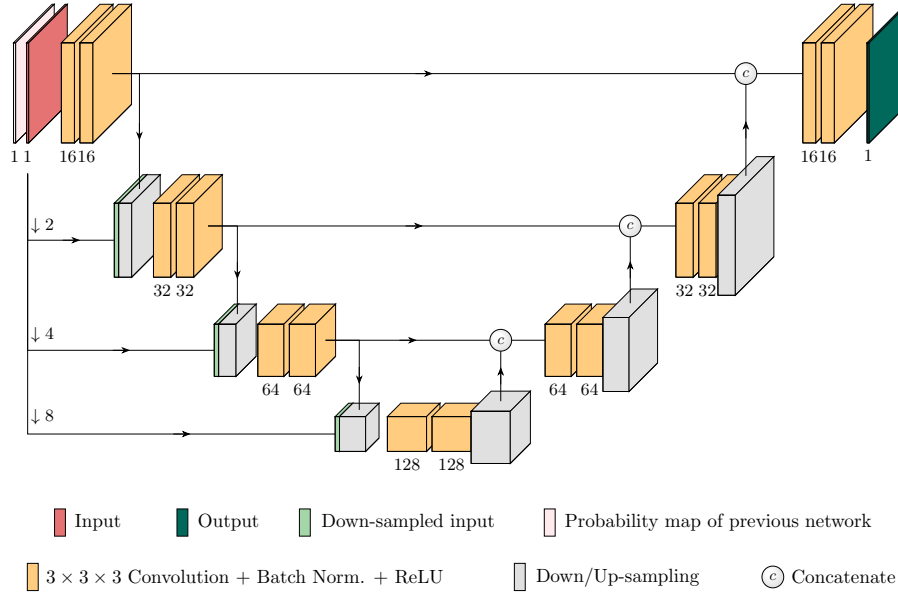- Down/Up-sampling
- ⓒ Concatenate

Fig. 3: The W-Net and T-Net architecture. The figure is simplified to the single-slice case, however the convolutional layers were 3D.

2.3   Squeeze-and-Excitation Block

After each convolution and concatenation block, we added an Squeeze-and-Excitation block (SEB) as developed by Hu *et al.* (2018) [4]. Including SEB is a computationally lightweight but effective method for incorporating channel-wise interdependencies, and has shown to improve network performance by a significant margin [12].

3   Experiments

The proposed method was implemented in the Keras library[1] using TensorFlow[2] as the backend. This research was conducted using the resources of the High Performance Computing Center North (HPC2N), where the experiments were run on NVIDIA Tesla V100 GPUs.

3.1   Material

The data set provided is pre- and postoperative, arterial phase abdominal CT data from 300 randomly selected kidney cancer patients that underwent radical nephrectomy at the University of Minnesota Medical Center between 2010

---

[1] https://keras.io
[2] https://tensorflow.org

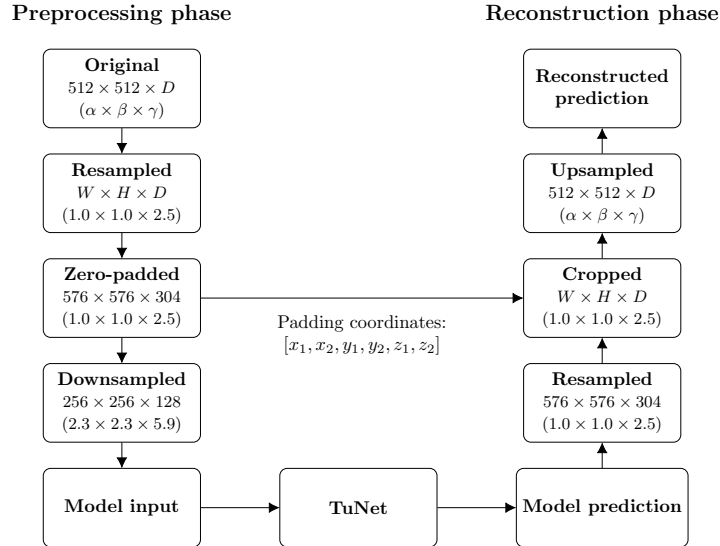**Preprocessing phase**   **Reconstruction phase**

Fig. 4: Preprocessing and reconstruction pipeline method as applied on the KiTS19 data set. Given are the resolutions and in parentheses the voxel dimensions in mm. W, H, and D each denote that the volume shape is varied in width, height, and/or depth, respectively. α, β, and γ each denote that the voxel spacing is varied in width, height, and/or depth, respectively. $[x_1, x_2, y_1, y_2, z_1, z_2]$ denote the number of pixels padded above and below in the x, y and z directions, respectively.

and 2018 [3]. Medical students annotated under supervision the contours of the whole kidney including any tumors and cysts, and contours of only the tumor component excluding all kidney tissue. Afterwards, voxels with a radiodensity of less than -30 HU were excluded from the kidney contours, as they were most likely perinephric fat.

Of the complete data set, 210 samples were initially released for training and validation on March 15, 2019. The remaining 90 samples were released without ground truth on July 15. The submission deadline for the predictions on the latter set was July 29.

3.2   Training Details

In addition to the experiments with single patch sizes, we also additionally denoised the data using a $3 \times 3 \times 3$ median filter. This yielded a total of four different experiments.

The loss function that was used contained two terms,

$$\mathcal{L} = \mathcal{L}_{whole} + \mathcal{L}_{tumor}, \tag{1}$$

where $\mathcal{L}_{whole}$ and $\mathcal{L}_{tumor}$ denote the soft dice loss of the whole (kidney + tumor) and tumor regions, respectively. Here, the soft dice loss for any structure was defined as

$$\mathcal{L}_{dice} = \frac{-2\sum_i u_i v_i}{\sum_i u_i + \sum_i v_i + \epsilon}, \tag{2}$$

where u_i is the SOFTMAX output produced by either the W-Net or the T-Net, v_i is a one-hot encoding of the ground truth segmentation map, and ϵ is a small constant added to avoid division by zero.

For evaluation of the segmentation performance, we employed the DSC which is defined as

$$D(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}, \tag{3}$$

where X and Y denote the output segmentation and its corresponding ground truth, respectively. We used five-fold cross-validation, and computed the mean DSC over the folds.

We used the Adam optimizer [5] with a learning rate of $1 \cdot 10^{-4}$ and momentum parameters of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We also used L_2 regularization (with parameter $1 \cdot 10^{-5}$), applied to the kernel weight matrix, for all convolutional layers to cope with overfitting. The activation function of the final layer was the sigmoid function.

We trained 20 models for 200 epochs (see Table 1). To prevent overfitting, we selected a patience period by dropping the learning rate by a factor of 0.2 if the validation loss did not improve over 6 epochs. In addition, the training process was stopped if the validation loss did not improve after 15 epochs. The training time was approximately 27 hours per model.

3.3 Patch Extraction

We employed a patch-based approach in which we set up the experiments with a patch shape of $128 \times 128 \times 128$. To extract more patches from each image, we allowed a patch overlap of $96 \times 96 \times 96$. The size of the patch shape was large enough to contain each kidney, when centered.

3.4 Preprocessing and Augmentation

All images in the data set had 512×512 resolution with varying slice counts in the axial direction. The voxel sizes also varied between patients. A preprocessing pipeline (see Figure 4) was set up to transform the data to uniform resolutions and voxel sizes.

First, the data were resampled to equal voxel size of $256 \times 256 \times 128$. The volumes were then zero-padded to the size of the single largest volume after resampling. In order to increase processing speed and lower the memory consumption, the volumes were thereafter downsampled to size $256 \times 256 \times 128$. As a final step, the images were normalized to zero-mean and unit-variance.

In order to augment the data set size and diversify the data, we employed simple on-the-fly data augmentation by randomly rotating the images within a range of -1 to 1 degrees.

3.5   Post-processing

Algorithm 1 shows our proposed post-processing pipeline. Each 3D probability map for the whole region was binarized using a threshold $T = 0.5$, which in case of an empty mask, was lowered to $T = 0.1$.

The corresponding probability map for the tumor region was first binarized using the threshold $T = 0.5$ which was lowered based on the following three cases: If the returned mask is empty, T is lowered to 0.1. If the returned mask is not empty but it contains less than 100 pixels, the lowered threshold is $T = 0.2$, and in any other case T is lowered to 0.3 to include more positives. The optimal value for T was tuned using the training data set.

The binary masks for the whole and tumor regions were then refined using a "region-growing" technique: given two thresholds. First, two binarized maps, A and B, were generated using the larger and smaller threshold, respectively. Second, the binary map, B, was labelled into different unconnected regions. Third, if a segmented region was connected to the binary map, A, it was then merged into A. The output of the refining step is the expanded region of A.

The refined binary masks were finally merged with the output of L-Net to produce a multi-class prediction map.

3.6   Ensembling

The median denoising and SEB led to four experiments: the baseline method (TuNet), the baseline on median denoised data (TuNet + median), the baseline with Squeeze-and-Excitation blocks (TuNet + SEB), and the baseline with Squeeze-and-Excitation blocks on median denoised data (TuNet + SEB + median). As a fifth experiment, an average ensemble technique was implemented, using the output probability maps of the above mentioned models, as

$$p_k = \frac{1}{M} \sum_{m=1}^{M} f_k, \tag{4}$$

where $p_k \in \mathbb{R}^{|K|}$ and $f_k \in \mathbb{R}^{|K|}$ denote the final probability of label k and the probability of label k generated by model m at an arbitrary voxel, respectively. Here, K is the set of tumor labels, *i.e.* $k = 1, 2$.

Each model was trained five independent times, their ensemble network (Ensemble) is therefore using a total of 20 models. As a last experiment, this ensemble model was used with an extra post-processing step (Ensemble + post-process) described in Algorithm 1.

Algorithm 1: Post-process

   Result: Post-processed prediction map

1 set global threshold $T = 0.5$;

2 binarize probability maps of whole and tumor regions based on T;

3 **if** *number of non-zeros voxels of binarized whole region equals 0* **then**

4 | set $T_{whole} = 0.1$;

5 | binarize probability map of whole region based on T_{whole};

6 **else**

7 | set $T_{whole} = 0.4$;

8 | refine whole region based on T and T_{whole};

9 **end**

10 **if** *number of non-zeros voxels of binarized tumor region equals 0* **then**

11 | set $T_{tumor} = 0.1$;

12 | binarize probability map of tumor region based on T_{tumor};

13 | **if** *number of non-zeros voxels of binarized tumor region equals 0* **then**

14 | | set tumor equal to whole region;

15 | **else**

16 | **end**

17 | refine tumor region based on T and T_{tumor};

18 **else**

19 | **if** *number of non-zeros voxels of binarized tumor region smaller than 100* **then**

20 | | set $T_{tumor} = 0.2$;

21 | | refine tumor region based on T and T_{tumor};

22 | **else**

23 | | set $T_{tumor} = 0.3$;

24 | | refine tumor region based on T and T_{tumor};

25 | **end**

26 **end**

27 merge whole and tumor regions to produce multi-class prediction map;

28 multiply prediction map with output of L-Net to generate final output;

---

4 Results and Discussion

Evaluations are collected below for the results of the vanilla form of U-Nets [11], for TuNet, its three variations: TuNet + median, TuNet + SEB, and TuNet + SEB + median, and the ensemble models: Ensemble and Ensemble + post-process. They underwent evaluations based on the DSC metric, the best variation was then also evaluated visually, see Figure 5.

4.1 Quantitative analysis

Table 1 shows the mean DSC score and standard deviation (SD) computed from the five-folds of cross-validation on the training set. The baseline method produced acceptable scores for the whole region segmentation, while the significant impairment of results for tumor segmentation are important to point out. They

Table 1: Mean DSC and SD computed from the five-folds of cross-validation on the training set. TuNet, median, SEB denote the baseline, with median denoising and with Squeeze-and-Excitation block, respectively. Note that the results were reported on the resampled output (size of $256 \times 256 \times 128$, see Figure 4).

Model	Sørensen-Dice coefficient		
	Kidney + Tumor	Tumor	Mean
U-Net 2D	0.776 (0.009)	0.174 (0.036)	0.475 (0.017)
U-Net 2D + median	0.781 (0.016)	0.172 (0.030)	0.476 (0.020)
U-Net 3D	0.793 (0.038)	0.154 (0.019)	0.474 (0.016)
U-Net 3D + median	0.809 (0.075)	0.192 (0.056)	0.501 (0.065)
TuNet	0.857 (0.029)	0.308 (0.034)	0.582 (0.030)
TuNet + median	0.878 (0.010)	0.346 (0.053)	0.612 (0.031)
TuNet + SEB	0.868 (0.020)	0.302 (0.035)	0.585 (0.023)
TuNet + SEB + median	0.864 (0.033)	0.292 (0.017)	0.578 (0.023)
Ensemble	0.894 (0.014)	0.331 (0.025)	0.612 (0.026)
Ensemble + post-process	**0.902 (0.014)**	**0.408 (0.031)**	**0.655 (0.019)**

are possibly due to the scarcity of tumors in the data set and because of their spatial inconsistency, which the proposed network could not invalidate.

The SEB variations improved the DSC scores over the baseline method for the whole region while impairing the tumor segmentation, leading to only a slight effect on the mean score, whether using median denoised data or not.

However, the baseline method using median denoised data significantly improved the DSC scores for both the whole and the tumor region, and although the SD for the tumor masks also increased, the improvements and effects of median denoising are still clear and important.

The Ensemble model reached the same mean score as the best of the other variations, *i.e.* TuNet + median. The post-processing step significantly improved the Ensemble model, making it the best-performing model of all models.

4.2   Qualitative analysis

Figure 5 demonstrates the inter-patient diversity of the structures in this data set, where the results are produced using the Ensemble + post-process model. The renal tumors can be present in left, right, or both kidneys, and their size in the images can vary from several cubic centimeters to multiple times the volume of the kidney it is attached to. The intratumoral radiodensity can also be either very homogeneous or very heterogeneous. This results in a challenging segmentation task. The morphology of the kidneys is however relatively comparable across patients and can therefore be annotated with a fairly consistent and acceptable accuracy.
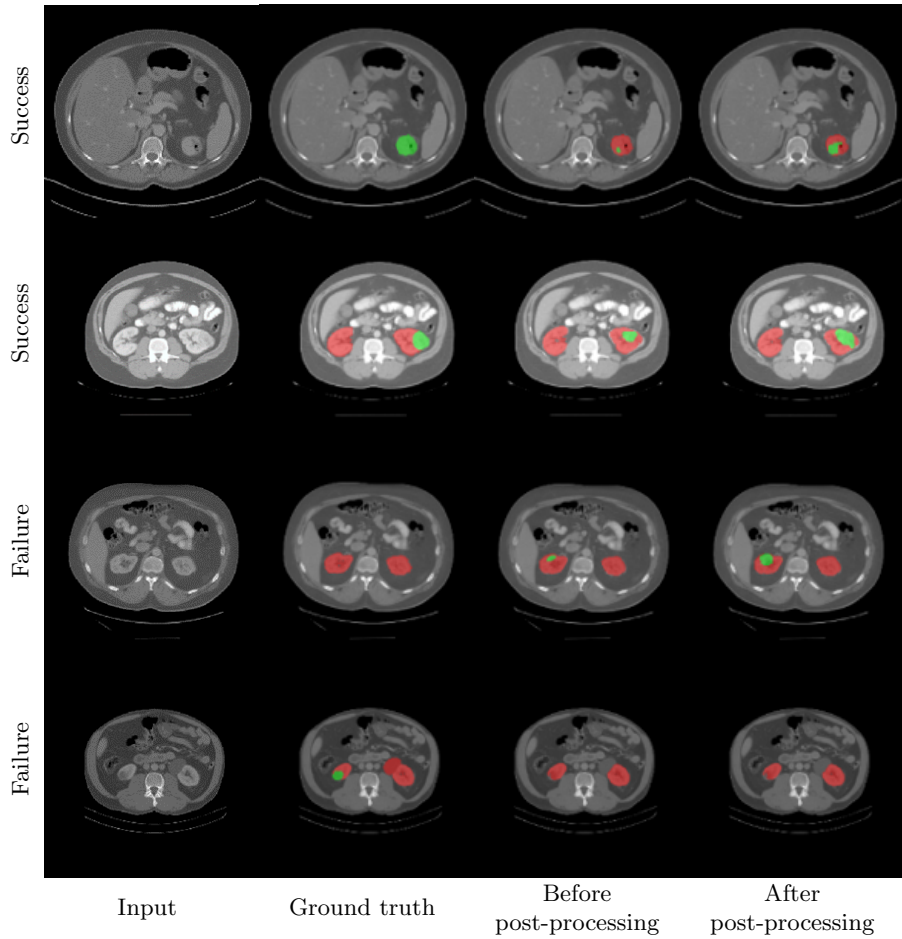
Fig. 5: A comparison of the ground truth masks and the results of Ensemble + post-process. The four examples show the input CT images, the ground truth masks and the results of the proposed method. For both cases the masks show the kidney+tumor region in red while the tumors are in green.

5   Conclusion

The KiTS19 Challenge is clearly justified given the difficulty of segmenting the images in the data set. Initially the provided data set was in itself anticipated to be insufficient for excelling results for this work. For training the proposed TuNet architecture, we used only a simple on-the-fly data augmentation, and still produced reasonable DSC scores for both the whole region and the tumor region compared to their corresponding models using a simple U-Net in both 2D and 3D.

Two expansions of the TuNet method were evaluated. Using SEB has only a slight effect on the mean scores, while median denoising the data leads to significant improvements. An ensemble model using these variations with a simple post-processing step outperformed all other models and reached impressive DSC scores.

References

1. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **68**(6), 394–424 (2018)
2. Capitanio, U., Bensalah, K., Bex, A., Boorjian, S.A., Bray, F., Coleman, J., Gore, J.L., Sun, M., Wood, C., Russo, P.: Epidemiology of renal cell carcinoma. European urology (2018)
3. Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., Dean, J., Tradewell, M., Shah, A., Tejpaul, R., Edgerton, Z., Peterson, M., Raza, S., Regmi, S., Papanikolopoulos, N., Weight, C.: The KiTS19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes (2019)
4. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
5. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
6. Kitrungrotsakul, T., Iwamoto, Y., Han, X.H., Takemoto, S., Yokota, H., Ipponjima, S., Nemoto, T., Wei, X., Chen, Y.W.: A cascade of CNN and LSTM network with 3D anchors for mitotic cell detection in 4D microscopic image. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1239–1243. IEEE (2019)
7. Kopp, R.P., Liss, M.A., Mehrazin, R., Wang, S., Lee, H.J., Jabaji, R., Mirheydar, H.S., Gillis, K., Patel, N., Palazzi, K.L., et al.: Analysis of renal functional outcomes after radical or partial nephrectomy for renal masses 7 cm using the renal score. Urology **86**(2), 312–320 (2015)
8. Kutikov, A., Uzzo, R.G.: The renal nephrometry score: a comprehensive standardized system for quantitating renal tumor size, location and depth. The Journal of urology **182**(3), 844–853 (2009)
9. Mehrazin, R., Palazzi, K.L., Kopp, R.P., Colangelo, C.J., Stroup, S.P., Masterson, J.H., Liss, M.A., Cohen, S.A., Jabaji, R., Park, S.K., et al.: Impact of tumour morphology on renal function decline after partial nephrectomy. BJU international **111**(8), E374–E382 (2013)
10. Motzer, R.J., Bander, N.H., Nanus, D.M.: Renal-cell carcinoma. New England Journal of Medicine **335**(12), 865–875 (1996)
11. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
12. Roy, A.G., Siddiqui, S., Pölsterl, S., Navab, N., Wachinger, C.: 'squeeze & excite' guided few-shot segmentation of volumetric images. CoRR **abs/1902.01314** (2019)

13. Wahlgren, T., Harmenberg, U., Sandström, P., Lundstam, S., Kowalski, J., Jakobsson, M., Sandin, R., Ljungberg, B.: Treatment and overall survival in renal cell carcinoma: a swedish population-based study (2000–2008). British journal of cancer **108**(7), 1541 (2013)
14. Wang, G., Li, W., Ourselin, S., Vercauteren, T.: Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In: International MICCAI Brainlesion Workshop. pp. 178–190. Springer (2017)
15. Yang, G., Li, G., Pan, T., Kong, Y., Wu, J., Shu, H., Luo, L., Dillenseger, J.L., Coatrieux, J.L., Tang, L., et al.: Automatic segmentation of kidney and renal tumor in CT images based on 3D fully convolutional neural network with pyramid pooling module. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 3790–3795. IEEE (2018)
16. Yu, Q., Shi, Y., Sun, J., Gao, Y., Zhu, J., Dai, Y.: Crossbar-Net: A novel convolutional neural network for kidney tumor segmentation in CT images. IEEE Transactions on Image Processing (2019)