On the influence of Dice loss function in multi-class organ segmentation of abdominal CT using 3D fully convolutional networks

7 authors, including:

Holger Reinhard Roth
Nagoya University
75 PUBLICATIONS **967** CITATIONS

Hirohisa Oda
Nagoya University
20 PUBLICATIONS **10** CITATIONS

Masahiro Oda
Nagoya University
101 PUBLICATIONS **239** CITATIONS

Kazunari Misawa
Aichi Cancer Center, Japan, Nagoya
111 PUBLICATIONS **737** CITATIONS

Some of the authors of this publication are also working on these related projects:

Project   computer aided diagnosis and surgery View project

Project   deeply exciting stuff! View project

On the influence of Dice loss function in multi-class organ segmentation of abdominal CT using 3D fully convolutional networks

Chen SHEN[1], Holger R. ROTH[1], Hirohisa ODA[2], Masahiro ODA[1], Yuichiro HAYASHI[1], Kazunari MISAWA[3], and Kensaku MORI[1]

[1]Graduate School of Informatics, Nagoya University
[2]Graduate School of Information Science, Nagoya University
[3]Aichi Cancer Center Hospital

Abstract

Deep learning-based methods achieved impressive results for the segmentation of medical images. With the development of 3D fully convolutional networks (FCNs), it has become feasible to produce improved results for multi-organ segmentation of 3D computed tomography (CT) images. The results of multi-organ segmentation using deep learning-based methods not only depend on the choice of networks architecture, but also strongly rely on the choice of loss function. In this paper, we present a discussion on the influence of Dice-based loss functions for multi-class organ segmentation using a dataset of abdominal CT volumes. We investigated three different types of weighting the Dice loss functions based on class label frequencies (uniform, simple and square) and evaluate their influence on segmentation accuracies. Furthermore, we compared the influence of different initial learning rates. We achieved average Dice scores of 81.3%, 59.5% and 31.7% for uniform, simple and square types of weighting when the learning rate is 0.001, and 78.2%, 81.0% and 58.5% for each weighting when the learning rate is 0.01. Our experiments indicated a strong relationship between class balancing weights and initial learning rate in training.

Key words: multi-organ segmentation, Dice loss function, fully convolutional network, computed tomography

1 Introduction

Multi-class organ segmentation of abdominal computed tomography (CT) images is important for medical image analysis. Abdominal organs can have large individual differences due to the shape and size variations, which makes development of automated segmentation methods challenging. With the rapid development of medical image devices and the increase in routinely acquired images, fully automatic multi-organ segmentation of medical images becomes especially important. The segmentation results can be widely utilized in computer-aided diagnosis and computer-assisted surgery. Hence, much research [1][2][3] has focused on automated organ segmentation from CT volumes. However, achieving high segmentation accuracy is always challenging. One of the reasons is the low contrast to surrounding tissues which makes segmentations difficult. Failure of segmentation results in a reduction of diagnostic quality. Therefore, improving the segmentation accuracy is an active area of research. Recently, deep learning-based methods achieved impressive segmentation results on medical images. For example, convolutional neural networks (CNNs) make it easier to train models on a large datasets. Especially, 3D fully convolutional networks (FCNs) improved the accuracy of multi-organ segmentation from CT volumes. It is well-known that network architecture influences the result of segmentation. Furthermore, the segmentation accuracy relies on the choice of loss function. For this research, we developed a 3D FCN which can learn to automatically segment organs on CT volumes from a set of CT and labelled images, and discuss the influence of loss functions and initial learning

rates. The paper is organized as the follows. In Section 2, we describe the methods we utilized in detail. Section 3 gives the experiments and results, and we provide a discussion in Section 4.

2 Methods

2.1 Overview

FCNs have the ability to train models for segmentation from images in end-to-end fashion. With the development of FCNs, improved segmentation results have been reported[4]. In this paper, we utilize a 3D U-Net architecture for multi-organ segmentation, similar to the one proposed by Çiçek et al[5]. This network architecture is a type of 3D FCN designed for end-to-end image segmentation. We investigated three different types of weighting models for the Dice loss functions in order to evaluate their performances and compared segmentation accuracy of each organ for the different situations. In addition, we performed training on the same dataset with different initial learning rates μ, and compared how these parameters influence the segmentation results.

2.2 3D fully convolutional network

With the improvement of CNN architectures and GPU memory, we are able to train a larger number of annotated 3D medical CT volumes to improve the segmentation results. We utilize a training set of abdominal CT volumes and labels $S = (I_n, L_n), n = 1, ..., N$, where I_n represents CT volumes, L_n represents the ground truth label volumes. We define n as the index of volumes and N is the number of training volumes.

In this work, we utilize a 3D U-Net type FCN with constant input size that is trained by randomly cropping subvolumes from the training data. Hence, we can obtain a trained model, which can segment the full 3D volume through subvolume prediction in testing. We chose an input size of 646464 that allows to use mini-batch sizes of three subvolumes sampled from different training patients.

2.3 Dice loss function

The Dice similarity coefficient (DSC) measures the amount of agreement between two image regions[6]. It is widely used as a metric to evaluate the segmentation performance with the given ground truth in medical images. The DSC is defined in (1), we utilize | | to indicate the number of foreground voxels in the ground truth and segmentation images[7].

$$DSC = \frac{2|S \cap R|}{|S| + |R|} \tag{1}$$

where S is the segmentation result and R is the corresponding ground truth label. This function however is not differentiable and hence cannot directly be used as a loss function in deep learning. Hence, continuous versions of the Dice score have been proposed that allow differentiation and can be used as loss in optimization schemes based on stochastic gradient descent[8]:

$$\mathcal{L}_{DSC} = -\frac{2\sum_i^N s_i r_i}{\sum_i^N s_i + \sum_i^N r_i} \tag{2}$$

where s_i and r_i represent the continuous values of the *softmax* prediction map $\in [0, ..., 1]$ and the ground truth at each voxel i, respectively. Using the formulation in 2, we investigate three types of weighting based on class voxel frequencies in the whole training dataset. We defined three types of weighting factors W_u, W_s and W_q for *uniform*, *simple* and *square* weighting. The equations are as follows:

$$W_u = 1 \tag{3}$$

$$W_s = \frac{N}{L|R_l| + \epsilon} \tag{4}$$

(a) Initial learning rate $\mu = 0.001, W_u$ (b) Initial learning rate $\mu = 0.001, W_s$ (c) Initial learning rate $\mu = 0.001, W_q$

(d) Initial learning rate $\mu = 0.01, W_u$ (e) Initial learning rate $\mu = 0.01, W_s$ (f) Initial learning rate $\mu = 0.01, W_q$
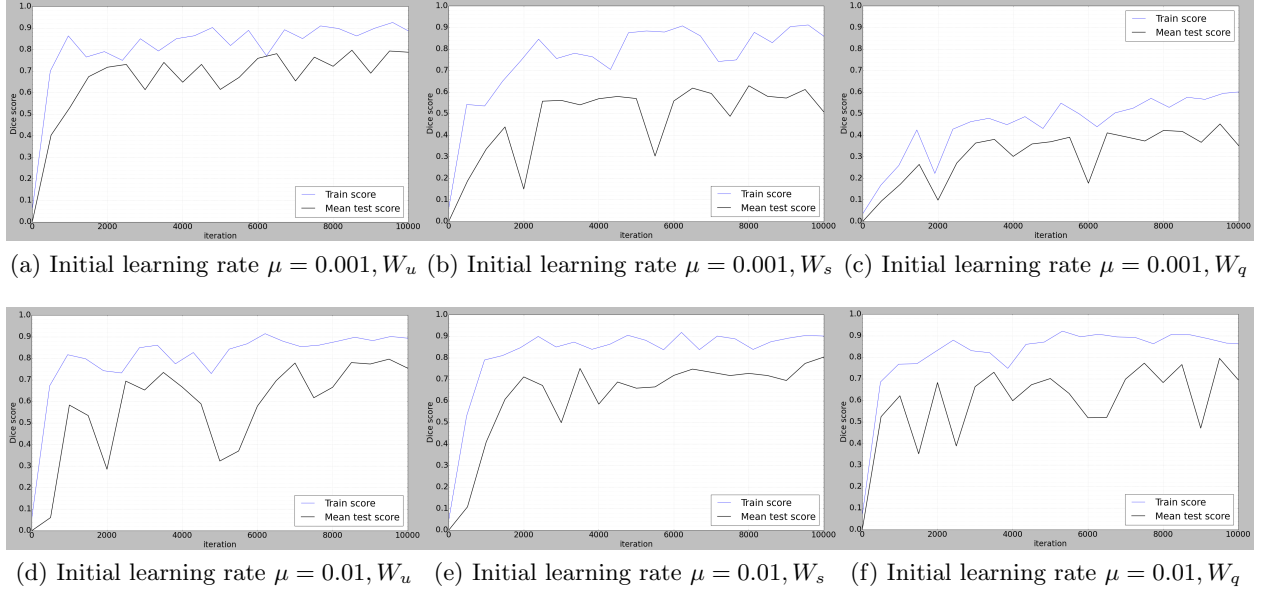
Figure 1: Learning curves for different models.

$$W_q = \frac{N}{L|R_l|^2 + \epsilon} \tag{5}$$

here, L is the number of labels and $|R_l|$ is the number voxels in class l. We set $\epsilon = 1$ in this experiment in order to avoid division by zero.

We calculate the multi-class Dice loss function using 6:

$$\mathcal{L}_{MD} = \frac{1}{L} \sum_l^L w_l L_{DSC} \tag{6}$$

with w_l indicating the weight for class l, computed from one of the weighting types W_u, W_s, W_q, .

3 Experiments and results

We used 377 abdominal clinical CT volumes in portal venous phase acquired as pre-operative scans for gastric surgery planning. Each CT volume contains 460-1177 slices of 512×512 pixels. We down-sampled the volumes with a factor of four in all experiments. We evaluated our models using a random subset of 340 training and 37 testing patients. The number of labelled abdominal organs is seven, which includes the artery, portal vein, liver, spleen, stomach, gallbladder and pancreas. Our network produced eight prediction maps as output (including the seven organ classes plus the background). Ground truth was established manually using semi-automated segmentation tools like graph-cuts and region growing using the Pluto software[10].

We implement our models in Keras[11] using the TensorFlow backend. We compared the Dice similarity scores of seven organs and background of the three different weight types: *uniform*, *simple* and *square*. Furthermore, we compared different initial learning rates of $\mu_1 = 0.001$ and $\mu_2 = 0.01$ using Adam[9] as optimization method. Training of each model was performed for 10,000 iterations which takes about two days on a NVIDIA GeForce GTX 1080 GPU with 8GB memory. The segmentation results of the models with different weightings and learning rates are shown in Table 1. The average Dice score performances were 88.3% for *uniform*, 74.2% for *simple* and 44.0% for *square* weighting-when the initial learning rate μ_1=0.001. When the initial learning rate was μ_2=0.01, the Dice score for *uniform*, *simple* and *square* is 86.4%, 88.9% and 78.9%, respectively.

Table 1: Average Dice scores of different type of weighting schemes (*uniform, simple, square*). μ_1 and μ_2 is the value of initial learning rate. We define μ_1=0.001 and μ_2=0.01 in this experiment. The highest accuracies for each organ are written in **bold** text.

Dice type	$W_u, \mu_1 = 0.001$	$W_s, \mu_1 = 0.001$	$W_q, \mu_1 = 0.001$	$W_u, \mu_2 = 0.01$	$W_s, \mu_2 = 0.01$	$W_q, \mu_2 = 0.01$
background	**99.9%**	99.1%	79.2%	99.8%	99.8%	99.6%
artery	80.4%	63.6%	63.4%	79.5%	**89.2%**	72.4%
vein	78.6%	68.8%	69.1%	75.1%	**79.3%**	73.0%
liver	**96.5%**	73.4%	10.1%	95.6%	96.3%	87.7%
spleen	**94.7%**	71.6%	0%	92.0%	93.9%	91.6%
stomach	**96.3%**	64.1%	6.9%	93.9%	96.1%	82.4%
gallbladder	77.3%	74.7%	54.3%	73.2%	**80.8%**	54.2%
pancreas	82.7%	78.0%	69.1%	82.4%	**84.7%**	70.3%
AVG	88.3%	74.2%	44.0%	86.4%	**88.9%**	78.9%
MAX	**99.9%**	99.1%	79.2%	99.8%	99.8%	99.6%
MIN	77.3%	63.6%	0.0%	73.2%	**79.3%**	54.2%

Figure 1 depicts the learning curve for comparison of different weighting types. Figures 2, 3 and 4 show the segmentation results of each weighting scheme.

4 Discussion

The results of our experiments indicate that the weighting of the Dice loss function and the initial learning rate both affected the performance of multi-organ segmentation in CT volumes. For the type of Dice loss function, the results showed clearly in Table 1 that when the learning rate was 0.001, the *uniform* model performed the highest on average DSC of eight classes in the CT volumes. It achieved impressive result for liver, spleen and stomach with average DSC of 93.7%, 91.6% and 91.1%, respectively. For other organs, although the results were not the best, they were still acceptable. The *simple* and *square* models performed much worse than *uniform* on liver, spleen, and stomach segmentation. However, the *simple* model performed best when the learning rate was increased to 0.01. The artery, portal vein, and pancreas performed best on this case, and the DSC of liver, spleen, stomach and gallbladder are also higher. As for the *uniform* weights, it kept a stable accuracy.

For the learning rate, the results also can be seen from Table 1. As we increased the learning rate from 0.001 to 0.01, the performances of *simple* and *square* models improved markedly. Especially for the *simple* type, the average DSC was raised from 59.5% to 81.0%. Furthermore, the results on artery, portal vein, and pancreas showed the best with the *simple* model when the initial learning rate was 0.01. Also, the *square* model performed much better on segmenting the liver, spleen, and stomach. Therefore, we can conclude that a higher learning rate was beneficial for *simple* and *square* type models in multi-organ segmentation. We inferred that the DSC of the *square* model could be even higher after the same number of iterations, if the learning rate was set to 0.1.

Moreover, the learning curves showed in Figure 1 indicated that the training converges to stable results for the *uniform* weighting when the iteration is 10,000. However, for the *simple* and *square* weighting model, we can predict that the performance could continue to grow if training would be continued. We assume that a convergence would be achieved after 20,000 iterations.

Utilizing three types of weighting and two different learning rates, the relationship between weighting type and learning rate can be observed from these experiments. We assume that the iteration number will also influence the results, but might result in overfitting. Our experiments indicate that when introducing a class balancing weight, the initial learning rates and number of iterations have to be adjusted appropriately in order to achieve improvements in the segmentation accuracy.

(a) Ground truth of CT image   (b) Learning rate $\mu = 0.001$ (c) Learning rate $\mu = 0.01$

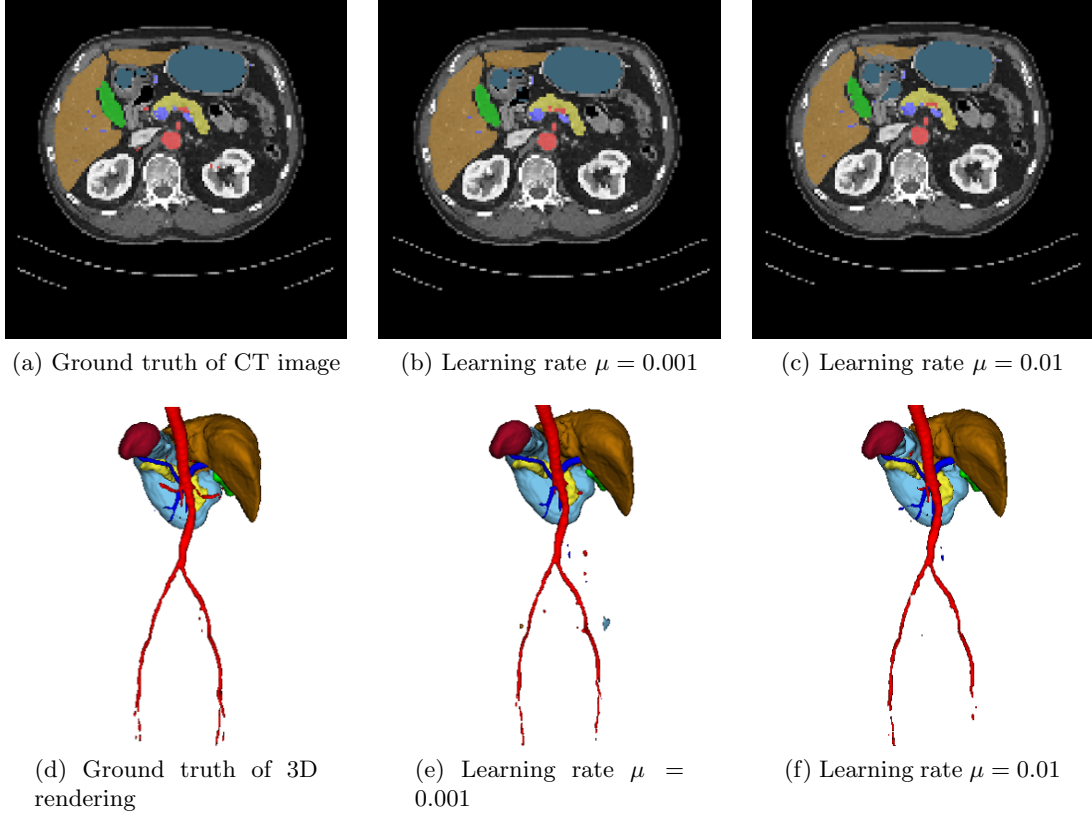(d) Ground truth of 3D rendering   (e) Learning rate $\mu = 0.001$ (f) Learning rate $\mu = 0.01$

Figure 2: Segmentation results for *uniform* type weight. The images in top row are CT axial slices and in the bottom row are 3D surface renderings of the segmentation results.

5   Conclusion

We employed three types of weighting models for a Dice-based loss function and evaluated the segmentation accuracy for multiple organs with two different initial learning rates. These different types of weight models shows the influence on the multi-organ segmentation in CT volumes using FCN. The results depict that the class balancing weights and initial learning rates influence the performance of multi-organ segmentation in CT volumes.

While we did not apply any data augmentation schemes, our results indicate no cases of strong overfitting, which points to a sufficiently large training dataset. Still, for future work, we can augment the original images and study how augmentation affects the segmentation accuracy. Moreover, we can match the best type of weighting and learning rate for single organs to achieve improvements in segmentation accuracy by combining the results from different models. With the availability of higher GPU memory and even larger datasets, the performance for automatic multi-organ segmentation is likely to increase.

6   Acknowledgments

(a) Ground truth of CT image   (b) Learning rate $\mu = 0.001$ (c) Learning rate $\mu = 0.01$

(d) Ground truth of 3D rendering   (e) Learning rate $\mu = 0.001$ (f) Learning rate $\mu = 0.01$
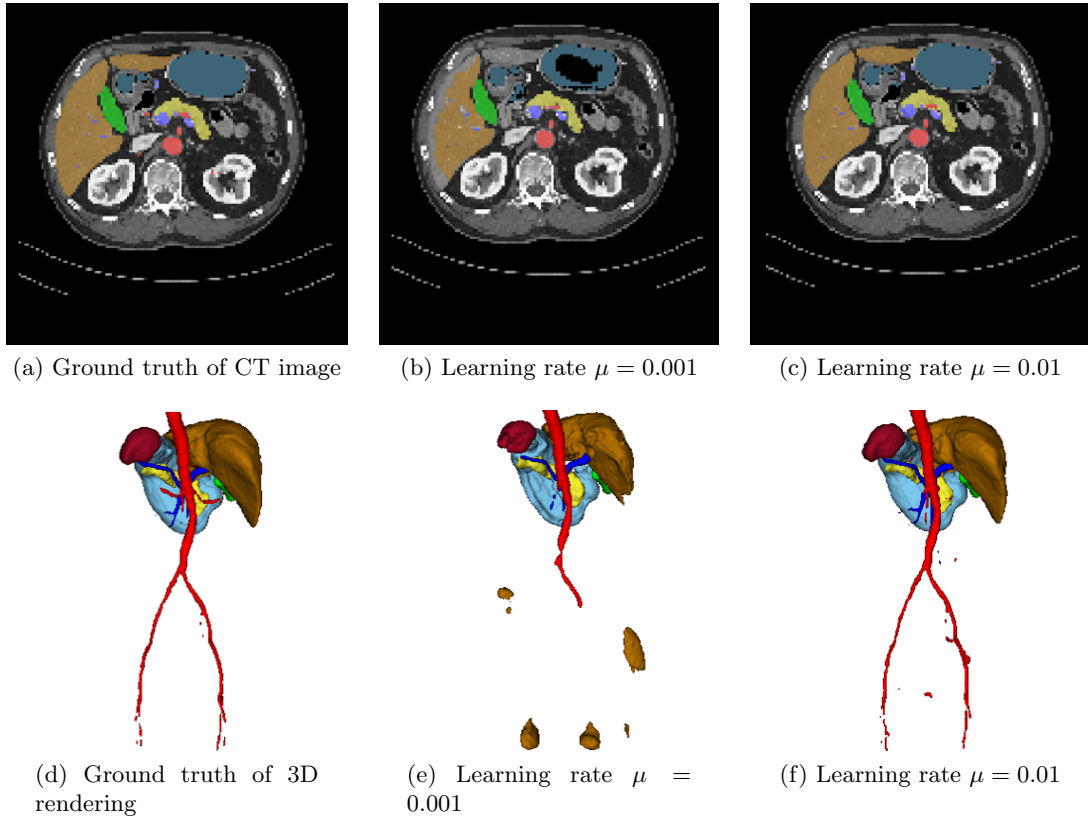
Figure 3: Segmentation results for *simple* type weight. The images in top row are CT axial slices and in the bottom row are 3D surface renderings of the segmentation results.

(a) Ground truth of CT image   (b) Learning rate $\mu = 0.001$ (c) Learning rate $\mu = 0.01$

(d) Ground truth of 3D rendering   (e) Learning rate $\mu = 0.001$ (f) Learning rate $\mu = 0.01$tab
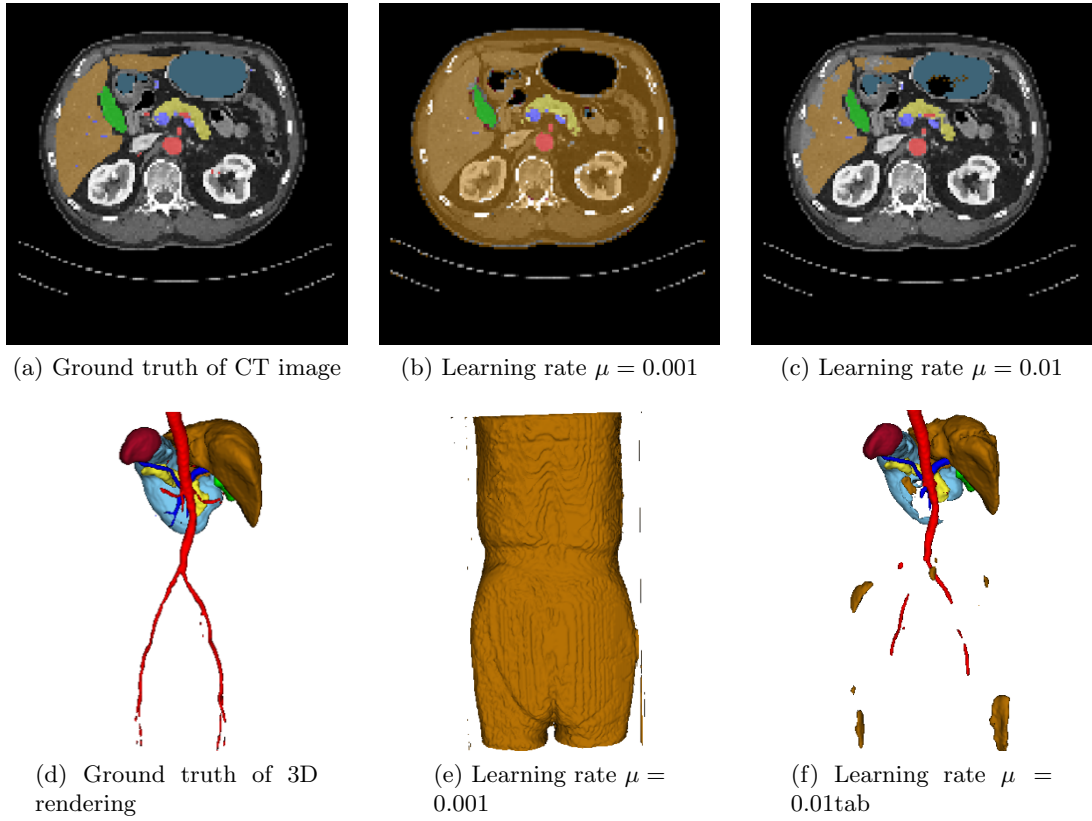
Figure 4: Segmentation results for *square* type weight. The images in top row are CT axial slices and in the bottom row are 3D surface renderings of the segmentation results.

References

[1] H. R. Roth, M. Oda, N. Shimizu, H. Oda, Y. Hayashi, T. Kitasaka, M. Fujiwara, K. Misawa, and K. MoriTowards dense volumetric pancreas segmentation in CT using 3D fully convolutional networks, SPIE. Medical Imaging, 2018.(accepted)

[2] H. R. Roth, M. Oda, Y. Hayashi, H. Oda, and K. Mori, Multi–organ segmentation in abdominal CT using 3D fully convolutional networks, International Journal of Computer Assisted Radiology and Surgery, Vol.12, Sup.1, pp.S55–S57, 2017.

[3] O. Ronneberger, P. Fischer, and T. Brox, U–Net: Convolutional Networks for Biomedical Image Segmentation, International Conference on Medical Image Computing and Computer-Assisted Intervention, vol.9351, pp.234–241, 2015

[4] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.3431–3440, 2015.

[5] Ç Özgün, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, 3D U–Net: Learning Dense Volumetric Segmentation from Sparse Annotation, MICCAI 2016, Part II, LNCS 9901, pp.424–432, 2016.

[6] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Cardoso, M. J. Generalized Dice overlap as a deep learning loss function for highly unbalanced segmentations, Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. DLMIA 2017, ML–CDS 2017, Lecture Notes in Computer Science, vol.10553, pp.240–248, 2017.

[7] A. A. Novikov, D. Lenis, D. Major, J. Hladuvka, M. Wimmer, and K. Bühler, Fully Convolutional Architectures for Multi-Class Segmentation in Chest Radiographs, Computing Research Repository, vol.abs/1701.08816, 2017.

[8] F. Milletari, N. Navab, and S. A. Ahmadi, V–net: Fully convolutional neural networks for volumetric medical image segmentation, 3D Vision (3DV), 2016 Fourth International Conference on. IEEE, 2016.

[9] D. P. Kingma, and J. L. Ba, Adam: A method for stochastic optimization, the 3rd International Conference for Learning Representations, 2015.

[10] Pluto, Computer aided diagnosis system for multiple organ and diseases, Mori laboratory Nagoya University, Japan, `http://pluto.newves.org`.

[11] Keras, The python Deep Learning library, `http://keras.io`.