Whole Heart Segmentation from CT images Using 3D U-Net architecture

Marija Habijan[1], Hrvoje Leventić[1], Irena Galić[1], Danilo Babin[2]
[1]Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, Osijek, Croatia
[2]imec-TELIN-IPI, Faculty of Engineering and Architecture, Ghent University, Belgium
marija.habijan@ferit.hr

***Abstract*—Recent studies have demonstrated the importance of neural networks in medical image processing and analysis. However, their great efficiency in segmentation tasks is highly dependent on the amount of training data. When these networks are used on small datasets, the process of data augmentation can be very significant. We propose a convolutional neural network approach for the whole heart segmentation which is based upon the 3D U-Net architecture and incorporates principle component analysis as an additional data augmentation technique. The network is trained end-to-end i.e. no pre-trained network is required. Evaluation of the proposed approach is performed on 20 3D CT images from MICCAI 2017 Multi-Modality Whole Heart Segmentation Challenge dataset, divided into 15 training and 5 validation images. Final segmentation results show a high Dice coefficient overlap to ground truth, indicating that the proposed approach is competitive to state-of-the-art. Additionally, we provide the discussion of the influence of different learning rates on the final segmentation results.**

***Keywords*—CT, data augmentation, heart segmentation, medical image segmentation, neural networks, volumetric segmentation**

I. Introduction

Cardiovascular diseases (CVDs) have been identified as one of the most common causes of death in the world [1]. Timely detection of CVDs allows physicians to set the correct treatment plan that can significantly improve and often save the patients' life. Continuous development and improvement of imaging techniques like echocardiography, computerized tomography (CT) and magnetic resonance makes such diagnosis possible. Using those techniques in combination with powerful medical image processing methods allows three-dimensional visual inspection of the heart and its substructures, consequently improving the patients' care. The main challenge of developing such methods is in the complexity of extracting useful anatomical information from a large amount of highly dimensional data. Often, the treatment for CVDs requires surgical interventions and an accurate heart segmentation represents a valuable tool for the pre-operative planning of such interventions. Accurate model of the heart can be obtained either through manual segmentation (an error prone, time-consuming and often subjective process), or by using specialized image processing methods for heart segmentation. The variations in anatomical structures of the heart make the development of automatic heart segmentation methods a challenging task. Various approaches are proposed to tackle this problem. Methods that separately segment specific anatomic structures are often based on deformable models [2]–[4]. Another commonly investigated type of approach is based on multi-class segmentation, often using atlas-based methods [5], [6]. Since cardiac images contain a huge amount of anatomical variability and unclear boundaries that are often caused by lacks in acquisition procedures, methods that incorporate prior knowledge and artificial neural networks have recently commonly used [7], [8].

Nevertheless, today's state-of-the-art methods in the field of medical image segmentation are diverse combinations of encoder and decoder neural network architecture. The main idea behind encoder-decoder architecture is in skip connections. The linkage of feature maps with high granularity from the decoder with fine-granular feature maps from the encoder provides segmentation masks with a precisely preserved details even on highly noisy background. The Long et al. [19] introduced fully convolutional neural network (FCN) where feature maps from the encoder are accepting and summing up-sampled feature maps. Similarly, U-Net [10] architecture links up-sampled features and adds convolutions between each sampling step. The importance of the skip connections is in their ability to restore the spatial resolution at the network's output. Zhou et al. [11] further improved the segmentation accuracy by introducing both dense and nested skip connections. This architecture reduces the gaps between encoder feature maps and decoder feature maps thus simplifying the optimization problem.

In this paper, we focus on the whole heart segmentation using a 3D U-Net architecture [12]. The primary contribution of our work is in the data augmentation process. Our augmentation process adds the principal component analysis (PCA) to the originally proposed on-the-fly elastic deformations and smooth dense field deformation. The paper is structured as follows. Used methods are described in Section 2. Implementation details and description of a dataset is presented in Section 3. Furthermore, quantitative analysis of conducted experiments as well as segmentation results are provided in Section 4. Finally, the conclusion is given in Section 5.

II. The Proposed Method

In this section we explain the proposed network architecture and the process of data augmentation – mainly the effects of principal component analysis (PCA) on the input and output of the network. The PCA-based data augmentation represents the main scientific contribution of this paper.

A. Network Architecture

The neural network architecture of the proposed method is based on the 3D U-Net architecture and consists of the two main parts: (1) a contracting encoder part whose main task is the analysis of the whole image; and (2) the consecutive expanding decoder that produces a full-resolution segmentation.

Each layer of the encoder have two $3 \times 3 \times 3$ convolutions followed by rectified linear unit (ReLu). After each of them, $2 \times 2 \times 2$ max pooling layer is placed. Similarly, the decoder part consist of two $2 \times 2 \times 2$ upconvolutions that are also followed by two $3 \times 3 \times 3$ convolutions and ReLu. The last, $1 \times 1 \times 1$, convolution layer reduces the output channels to match the number of labels which is in our case 7. Furthermore, we are avoiding bottlenecks in both, encoder and decoder part, by using doubled values of the channels before each max pooling [13]. Moreover, the network uses batch normalization before every ReLu for faster convergence. The learning process involves the generation of the dense volumetric segmentations while only requiring two-dimensional annotated slices for training. This is possible because of the weighted softmax loss function that sets weights of the unlabeled voxels to zero consequently allowing learning from only labeled ones. An illustration of the previously described 3D U-Net architecture is presented in Figure 1.

B. Data Augmentation with PCA

The process of data augmentation is extensively used to improve training performance due to limited size of the training data, since the artificial neural networks require huge amounts of annotated data for effective learning. Essentially, data augmentation process increases the size of the training dataset through a series of image transformations. This paper proposes a modification to the data augmentation process of the U-Net architecture by introducing an additional PCA-based image transformation.

We can represent the transformations of the data augmentation as a sequence of operations performed on the training data. Let GT be the set of all training samples, where each data sample consists of two 3D volumes with same dimensions: the input CT volume I with voxel gray values in range $[0, 4095]$ and the corresponding labeled CT volume L with voxel gray values in range $[0, 7]$ (labels).

Originally, 3D U-Net architecture implements data augmentation with rotation, scaling and smooth dense deformation field techniques. Let θ be the set of the three mentioned transformations. Transformations are iteratively applied to both an input I volume and a labeled L volume, as follows:

$$h : X \mapsto X, h \in \theta, X \in \{I, L\} \tag{1}$$

where transformation parameters for each $h \in \theta$ are chosen randomly on-the-fly.

Thus, the entire process of data augmentation itself can be described with the following mapping:

$$\theta : GT \mapsto AGT \tag{2}$$

where GT represents original training dataset and AGT represents the augmented (transformed) dataset of GT. The inflated training dataset is thus defined as:

$$GT' = GT \cup AGT, \tag{3}$$

meaning that GT' contains both the original training dataset GT and all the respective transformations defined with θ. This final inflated dataset is used to train the network.

In order to further inflate the training dataset, we extend the transformation set θ with an additional transformation that performs principal component analysis using the singular value decomposition. Let $\mathbf{v} \in \mathbb{Z}^3$ denote the voxel position in an input image. Let $P_{\mathbf{v}}$ denote a set of grayscale values representing the vector of principal components after performing the singular value decomposition of an input image X at position $\mathbf{v}$. Our proposed transformation h_{pca} is defined with the following mapping:

$$h_{pca} : X \mapsto X', X \in \{I, L\} \tag{4}$$

where X' denotes the resulting image after transformation. The proposed transformation modifies the grayscale values of every voxel $\mathbf{v}$ in the input image in a following manner:

$$X'(\mathbf{v}) = \frac{1}{s_p} P_v [\alpha_1 \lambda_1]^T, \tag{5}$$

where α_1 denote random variable drawn from a Gaussian with $mean = 0$ and $\sigma = 0.1$, λ_1 denote i^th eigenvalue corresponding to the eigenvector P and s_p denotes the scaling parameter initialized to $5e6$.

C. Adaptive Moment Optimization

In neural network problems, the optimization algorithms aim to find optimal weights while simultaneously minimizing error and maximizing accuracy. Instead of originally used stochastic gradient descent for network weights updating, we used Adaptive Moment Optimization (Adam) [14] which iteratively updates weights during training process. The weights updated with stochastic gradient descent provide a single, fixed learning rate during training. Contrary, the Adam maintains a learning rate for each network weight as the learning unfolds. Adam incorporates the strenghts of the two stochastic gradient descent extensions [15]; the root mean square propagation (RMSprop) and adaptive gradient algorithm (AdaGrad) consequently providing an optimization algorithm that can handle sparse gradients on noisy problems.

Update equations, for each weight w_j can be defined as:

$$a_t = \beta_1 * a_{t-1} - (1 - \beta_1) * g_t \tag{6}$$

$$b_t = \beta_2 * b_{t-1} - (1 - \beta_2) * g_t^2 \tag{7}$$

$$\Delta w_t = -\zeta \frac{a_t}{\sqrt{b_t + \epsilon}} * g_t \tag{8}$$

where β_1 and β_2 are hyperparameters, ζ represents the initial learning rate, g_t is the gradient at time t along w_j, a_t and b_t represent exponential average of the gradients and squares of gradients along w_j respectively.

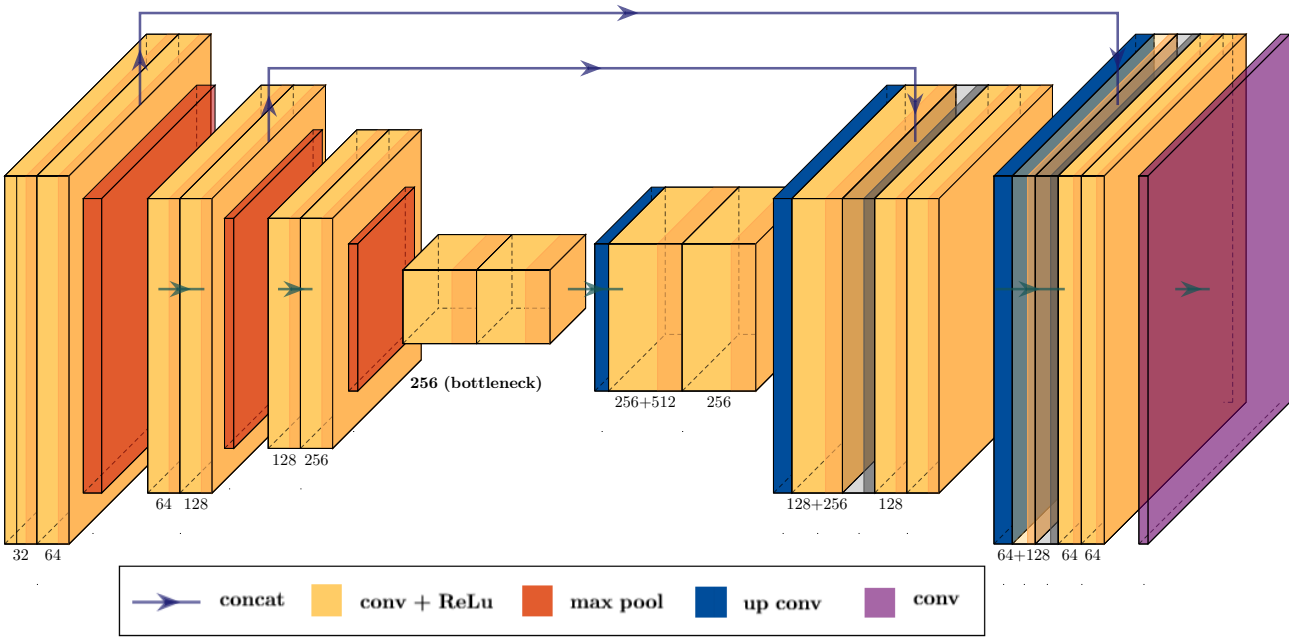

Fig. 1. Illustration of the 3D U-Net architecture

The exponential average and the squares of the gradient of each parameter is computed from Eq. 6 and Eq. 7 while decision of the learning step is calculated with Eq.8. Furthermore, common values of the hyperparameters are $\beta_1 = 0.9$, $\beta_2 = 0.99$ while ϵ is chosen to be $1e{-}10$ generally. In our implementation, we also use mentioned hyperparameters values.

III. Implementation Details

A. Dataset Description

Data provided in the MICCAI 2017 Multi-Modality Whole Heart Segmentation Challenge is acquired from the everyday clinical environment, using cardiac CT angiography and different scanner types. This resulted in images with varying voxel sizes and resolutions with the aim of providing imperfect training data to encourage the development of more robust algorithms. The slices were acquired in the axial view with the pixel resolution of 512×512. The average in-plane resolution is about 0.78×0.78 mm and the average slice thickness is 1.60 mm which leads to volumetric data consisting of 350 to 500 two-dimensional slices. Furthermore, the labeled data contains the whole heart i.e. all heart substructures from the upper abdomen to the aortic arch including the left ventricle (LV), the myocardium of the left ventricle (Myo), the left atrium (LA), the right ventricle (RV), the right atrium (RA) the pulmonary artery (PA) and the ascending aorta (AA).

B. Input Preprocessing

With the objective of simplifying the further process of the network training and reducing the computational time, all volumes used for training were preprocessed i.e. we normalize the intensity applying a following linear transformation:

$$out_p = (in_p - in_{min}) \cdot \frac{out_{max} - out_{min}}{in_{max} - in_{min}} + out_{min} \qquad (9)$$

where out_p represents wanted, output pixel values, in_p are all input pixel values while out_{min} and out_{max} are user defined parameters. We used values $out_{min} = 0$ and $out_{max} = 20000$. Therefore, all volumes are resampled, and the network takes three-dimensional volumes of the voxel size $144 \times 144 \times 144$ as input and returns the voxels of the same size at the final layer in x, y, and z directions respectively. Thus obtained volumes are then used and forwarded to the first convolutional layer at the network input.

C. Localization and Segmentation

The proposed approach consists of two architectures; one used for the whole heart localization while second used for its segmentation. It is important to emphasize that used ground-truth bounding box represents the whole heart rather than its specific substructures such as the left atrium or the pulmonary artery that are strongly connected together. The purpose of the segmentation network is to contain the only necessary information for the spatial regions of the interest from the whole volume consequently simplifying prediction process. The illustration of the used framework is shown in Figure 2.

IV. Experiments and Results

For the implementation, we use Keras and Tensorflow. Since training of the huge 3D networks requires high computational power, we use the cuDNN convolution layer implementation to increase memory efficiency. Data augmentation is done on-the-fly, which results in as many different artificial training images as the training iterations. We ran in total 120000 training

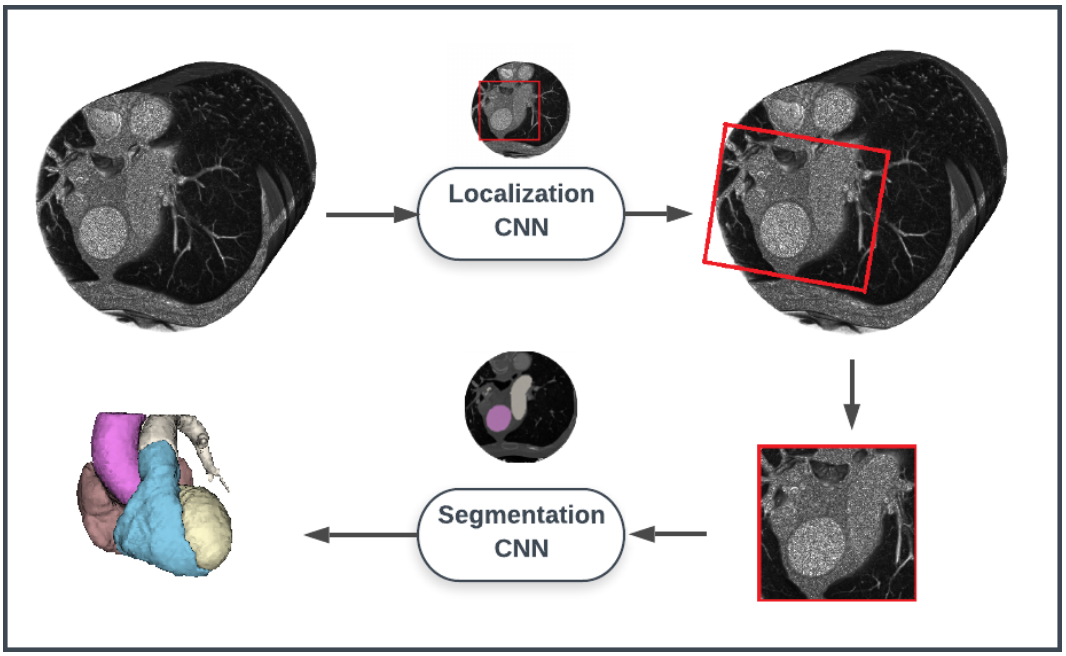

Fig. 2. The illustration of the proposed framework

TABLE I
DICE COEFFICIENT IN % FOR THE ORIGINAL 3D U-NET ARCHITECTURE TESTED ON FIVE VOLUMETRIC DATASETS.

Dataset	DSC of obtained results [%]							
	LV	RV	LA	RA	PA	Myo	Aorta	WH
1001	94.9	89.8	90.7	88.3	88.2	87.7	93.5	**90.5**
1002	93.3	89.5	91.3	88.1	85.5	87.3	91.7	**89.5**
1010	91.9	87.7	90.4	87.6	84.2	85.7	92.1	**88.5**
1019	87.9	83.6	86.4	83.3	79.9	81.5	87.7	**84.3**
1020	90.6	86.8	88.7	84.6	82.7	81.6	89.9	**86.4**
Average	91.72	87.48	89.5	86.38	84.1	84.76	91.06	**87.8**

TABLE II
DICE COEFFICIENT IN % FOR THE 3D U-NET ARCHITECTURE WITH IMPROVED DATA AUGMENTATION TESTED ON FIVE VOLUMETRIC DATASETS

Dataset	DSC of obtained results [%]							
	LV	RV	LA	RA	PA	Myo	Aorta	WH
1001	95.5	90.7	91.5	88.8	89.7	88.4	94.8	**91.3**
1002	94.6	89.9	91.8	88.9	86.9	88.3	93.8	**90.6**
1010	93.2	89.5	91.1	88.1	84.5	87.7	93.5	**89.6**
1019	89.9	85.1	87.7	83.5	80.1	81.6	89.7	**85.3**
1020	91.2	88.4	90.4	87.6	84.5	83.2	92.1	**88.2**
Average	92.88	88.72	90.5	85.14	85.14	85.84	92.78	**89.00**

iterations, simultaneously on two NVidia Geforce Titan V GPUs, which took approximately 25 hours. Furthermore, the segmentation of the new dataset took on average only 6.8 seconds on the same hardware.

A. Results

For the evaluation purposes, we use 20 CT volumes with the corresponding manually labeled ground-truths used in the MICCAI 2017 Multi-Modality Whole Heart Segmentation Challenge. We divided data into two sets. The training set consist of 15 CT volumes, while the validation set consist of the remaining 5 CT volumes. The quantitatively model performance during training is measured with the dice coefficient. The similarity of the predicted segmentation and the ground-truth label is calculated with the following formula:

$$dice(Y, Z) = \frac{2 \times |Y \cap Z|}{|Y| + |Z|} \tag{10}$$

In this manner, we obtained an overall average score for the whole heart segmentation of 89%. The table-like representation of the all obtained results for the original 3D U-Net architecture is shown in Table I while the original 3D U-Net architecture with data augmentation improvement is shown in Table II. An example of the best-segmented dataset, precisely dataset 1001, is shown in Figure 3.

We compared our result of the proposed method to the three similar whole heart segmentation approaches that use convolutional neural-networks. Payer et al. [17] obtained average dice score of 88.9% by combining landmark localization with a U-Net like CNN using heatmap regression and SpatialConfigura-

TABLE III
DIFFERENT INITIAL LEARNING RATES AND OBTAINED HEART SEGMENTATION RESULTS.

Initial Learning Rate	**DSC of obtained results [%]**								
	Dataset	LV	RV	LA	RA	PA	Myo	Aorta	WH
0.0001	1001	91.5	87.2	88.1	86.1	85.1	86.4	91.2	87.4
	1002	90.2	86.3	87.9	85.9	83.3	85.1	90.8	86.8
	1010	89.6	85.6	88.1	85.2	81.4	84.2	89.7	85.9
	1019	86.4	82.3	84.8	85.1	78.1	78.2	86.5	85.1
	1020	87.1	85.1	86.5	79.8	82.4	80.3	88.9	86.7
	Average	**88.96**	**85.3**	**87.08**	**84.42**	**82.06**	**82.84**	**89.96**	**86.38**
0.001	1001	93.7	89.1	90.3	87.4	87.3	87.2	93.5	90.1
	1002	93.2	88.0	90.8	86.3	84.2	86.9	92.5	88.5
	1010	91.5	87.3	90.5	87.3	82.9	85.8	91.3	87.3
	1019	87.3	83.5	86.3	81.6	79.5	79.9	87.9	86.1
	1020	89.7	86.7	88.8	86.1	83.2	81.2	90.3	87.1
	Average	**91.08**	**86.92**	**89.34**	**85.74**	**83.42**	**84.2**	**91.1**	**87.82**
0.005	1001	95.5	90.7	91.5	88.8	89.7	88.4	94.8	91.3
	1002	94.6	89.9	91.8	88.9	86.9	88.3	93.8	90.6
	1010	93.2	89.5	91.1	88.1	84.5	87.7	93.5	89.6
	1019	89.9	85.1	87.7	83.5	80.1	81.6	89.7	85.3
	1020	91.2	88.4	90.4	87.6	84.5	83.2	92.1	88.2
	Average	**92.88**	**88.72**	**90.5**	**85.14**	**85.14**	**85.84**	**92.78**	**89.00**

tion-Net architecture [18]. Wang et al. [19] developed a framework consisting of the 2.5D segmentation with orthogonal 2D U-nets, shape context estimation and refining segmentation with U-net and shape context. Furthermore, Xu, Wu and Feng [20] used the combination of Faster R-CNN and U-net network obtaining the overall average dice score of the $85,9\%$. It is important to point out that these three approaches, as well as ours, used the same validation dataset. This gives a reliable segmentation comparison that indicates competitiveness of our method to the state-of-the-art. Furthermore, for experimental purposes, we performed training with different initial learning rates and found the optimal learning rate of 0.005 to give best segmentation results as shown in Table III.

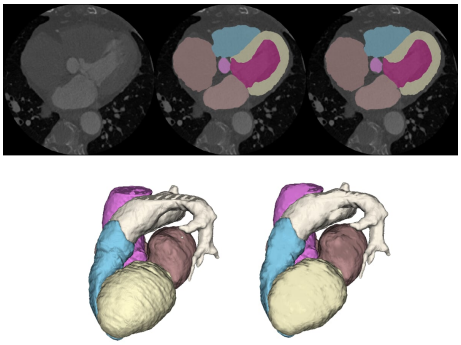

Fig. 3. The example of the segmentation result. From the upper left to the right: original 2D slice; 2D slice of the ground-truth; 2D slice of the segmentation result; 3D visualization of the ground-truth; 3D visualization of the predicted heart

V. CONCLUSION

We have presented an approach for automatic heart localization and segmentation from CT images. The proposed framework consists of two 3D U-Net neural network architectures, first used for localization of the bounding box around the heart and second used for the segmentation. The results were evaluated on the five CT volumes from the MICCAI 2017 Multi-Modality Whole Heart Segmentation challenge. We achieved an average dice score of 89% with additionally improved data augmentation process. Therefore, this approach offers a competitive and accurate segmentation method for the highly variable structures of the heart.

ACKNOWLEDGMENT

This work has been supported in part by Croatian Science Foundation under the project UIP-2017-05-4968.

REFERENCES

[1] HeartNet Connection, European Cardiovascular Disease Statistics 2017. http://www.ehnheart.org/cvd-statistics.html , 2017 (accessed 18 January 2019)

[2] Olivier Ecabert et al. “Automatic Model-Based Segmentation of the Heart in CT Images.” In: IEEE Transactions on Medical Imaging.(2008), vol. 27,pp. 1189–1201, doi: 10.1109/TMI.2008.918330

[3] O. Ecabert et al. ”Towards automatic full heart segmentation in computed tomography images.” In: Computers in Cardiology. (2005), doi: 10.1109/CIC.2005.1588077

[4] Xuan Zhao, Yao Wang, and Gabor Jozsef. “Robust shape-constrained active contour for whole heart segmentation in 3-D CT images for radiotherapy planning.” In: 2014 IEEE International Conference on Image Processing (ICIP). (2014), pp. 1–5, doi: 10.1109/ICIP.2014.7024999

[5] Zhuang, X., Rhode, K., Razavi, R., Hawkes, D.J., Ourselin, S.: A Registration-Based Propagation Framework for Automatic Whole Heart Segmentation of Cardiac MRI. IEEE Transactions on Medical Imaging, 29 (9): 1612-1625, 2010

[6] Xiahai Zhuang and Juan Shen: Multi- scale patch and multi-modality atlases for whole heart segmentation of MRI, Medical Image Analysis, vol.31, pp.77 - 87, 2016

[7] O. Oktay et al.”Anatomically Constrained Neural Networks (ACNNs): Application to Cardiac Image Enhancement and Segmentation” In: IEEE Transactions on Medical Imaging. (2017), vol. 37, no. 2, pp. 384-395, doi: 10.1109/TMI.2017.2743464

[8] O. Oktay et al."Anatomically Constrained Neural Networks (ACNNs): Application to Cardiac Image Enhancement and Segmentation" In: IEEE Transactions on Medical Imaging. (2018), vol. 37, no. 2, pp. 384-395, doi: 10.1109/TMI.2017.2743464

[9] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015.

[10] Ronneberger, Olaf & Fischer, Philipp & Brox, Thomas. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation.

[11] Zhou, Zongwei & Mahfuzur Rahman Siddiquee, Md & Tajbakhsh, Nima & Liang, Jianming. (2018). UNet++: A Nested U-Net Architecture for Medical Image Segmentation.

[12] Çiçek, Özgün et al. "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation." MICCAI (2016).

[13] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. CoRR abs/1512.00567 (2015)

[14] Kingma, Diederik and Ba, Jimmy. (2014). Adam: A Method for Stochastic Optimization. International Conference on Learning Representations.

[15] Chen, Jinghui and Quanquan Gu. "Closing the Generalization Gap of Adaptive Gradient Methods in Training Deep Neural Networks." CoRR abs/1806.06763 (2018): n. pag.

[16] Xiahai Zhuang: Challenges and Methodologies of Fully Automatic Whole Heart Segmentation: A review. Journal of Healthcare Engineering 4 (3): 371-407, 2013

[17] Payer, Christian & Štern, Darko & Bischof, Horst & Urschler, Martin. (2018). Multi-label Whole Heart Segmentation Using CNNs and Anatomical Label Configurations. 10.1007/978-3-319-75541-0_20.

[18] Payer, Christian & Štern, Darko & Bischof, Horst & Urschler, Martin. (2016). Regressing Heatmaps for Multiple Landmark Localization Using CNNs. 10.1007/978-3-319-46723-8_27.

[19] Wang, Chunliang & Smedby, Orjan. (2018). Automatic Whole Heart Segmentation Using Deep Learning and Shape Context. 10.1007/978-3-319-75541-0_26.

[20] Xu, Zhanwei & Wu, Ziyi & Feng, Jianjiang. (2018). CFUN: Combining Faster R-CNN and U-net Network for Efficient Whole Heart Segmentation.