2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Comparison of three U-Net family architectures for left ventricular myocardial wall automatic segmentation

Grigoris I Grigoriadis, Maria Roumpi, Dimitrios Zaridis, Vasilis C Pezoulas, Aidonis Rammos, Nikolaos S. Tachos, Katerina K Naka, Dimitrios I. Fotiadis, *Fellow, IEEE*

***Abstract*— Left ventricular (LV) segmentation is an important process which can provide quantitative clinical measurements such as volume, wall thickness and ejection fraction. The development of an automatic LV segmentation procedure is a challenging and complicated task mainly due to the variation of the heart shape from patient to patient, especially for those with pathological and physiological changes. In this study, we focus on the implementation, evaluation and comparison of three different Deep Learning architectures of the U-Net family: the custom 2-D U-Net, the ResU-Net++ and the DenseU-Net, in order to segment the LV myocardial wall. Our approach was applied to cardiac CT datasets specifically derived from patients with hypertrophic cardiomyopathy. The results of the models demonstrated high performance in the segmentation process with minor losses. The model revealed a dice score for U-Net, Res-U-net++ and Dense U-Net, 0.81, 0.82 and 0.84, respectively.**

Keywords: Left ventricular segmentation, Deep Learning, U-Net, CT imaging, Cardiac image analysis.

I. Introduction

The automatic identification of hypertrophic cardiomyopathy (HCM) using cardiac medical images is an emerging and challenging field. The development of medical imaging technologies provides the capability of early diagnosis and detection of the disease. Computed tomography (CT) among other imaging techniques is preferred for the visualization of the heart left ventricle (LV) and the evaluation of cardiomyopathies [1]. Segmentation and delineation of the left ventricle is a crucial step for the quantification of the morphological and pathological changes, providing important clinical variables, such as ejection fraction, end systolic and diastolic volume, wall thickness, etc. However, for most of the imaging modalities used, the manual segmentation of the heart is labor-intensive and time-consuming for a single subject [2]. Thus, automating the segmentation is highly desirable as it can provide significant contribution both in the clinical and the bioengineering domain.

Lately, relevant studies in LV segmentation focus mostly on deep learning (DL) techniques, as their results provide high accuracy. Specifically, the cardiac segmentation studies, utilize methods based on convolutional neural networks (CNN). A widely used established architecture, the U-Net [3], is used for biomedical image segmentation. Specifically, *Tong et al.* [4] proposed a deeply supervised 3D U-Net for fully automatic whole heart segmentation. In the training stage, a 3D U-Net was developed to detect the heart and segment the region of interest (ROI), where the training dataset was artificially augmented and, finally, a refined 3D U-Net was trained. Several studies [5, 6] have been also conducted by combining a localization network for the detection of the heart with 3D fully convolutional networks (FCNs), which were applied to detect the segmentation ROI, allowing the network to be more effective by focusing on the relevant anatomical regions. These methods achieve better segmentation accuracy, mainly due to the smaller variations in the image intensity distribution across different CT scanners and better image quality [7].

A variety of methods rely on the volumetric information extracted by the heart that used to train CNNs in different views (axial, sagittal, coronal views) in a 2D manner. *Wang et al.* [8] trained three independent orthogonal CNNs in order to segment different planes. Particularly, they employed a U-Net architecture that detects the ROI of the heart and classifies the pixels into different substructures without losing the original resolution. Additionally, they also integrated into the proposed framework a shape context for the segmentation refinement, whereas *Mortazi et al.* [9] developed an adaptive fusion strategy to combine multiple outputs from different views, with high segmentation accuracy calculated by the Dice Similarity Coefficient (DSC).

Different DL approaches utilize different loss functions (focal loss, Dice loss, categorical cross-entropy), which are combined to address the class imbalance among different ventricular structures and improve the segmentation performance [10, 11]. More recently, Jun Guo et al. [12] developed a 3D deeply supervised U-Net, which incorporates attention gates (AGs) to focus on the myocardial boundary structures and segment left ventricular myocardium contours. The literature uses mostly the sagittal view of the cardiac MRI scans, as these views provide a clear depiction of the LV as the targeted region of interest (ROI) is a circle (Apical to base LV). The anatomical shape depicted on the CT is more complex and challenging than the MRI.

* This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777204. This paper reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

G.I. Grigoriadis, Maria Roumpi, Dimitrios Zaridis, Vasilis C Pezoulas, Nikolaos Tachos and Dimitrios I. Fotiadis are with the Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, GR 45110 Ioannina, Greece (e-mails: greg8grigoriadis@gmail.com, mroumpi89@gmail.com, dimzaridis@gmail.com, bpezoulas@gmail.com, ntachos@gmail.com.

Dimitrios.I. Fotiadis is with the Department of Biomedical Research, Institute of Molecular Biology and Biotechnology, FORTH, Ioannina, Greece and the Dept. of Materials Science and Engineering, Unit of Medical Technology and Intelligent Information Systems, University of Ioannina, GR 45110, Ioannina, Greece (e-mail: fotiadis@cc.uoi.gr).

K.K. Naka, Aidonis Rammos is with the Michaelidion Cardiac Center, 2nd Department of Cardiology in the Faculty of Medicine, School of Health Sciences, University of Ioannina, GR 45110 Ioannina, Greece (emails: anaka@uoi.gr, aidrammos@yahoo.gr).

The aim of this paper is to present an implementation and a comparison study of three state of the art deep neural networks of the U-Net family, which have demonstrated the best performance. To the best of our knowledge, they have not been utilized for LV segmentation from CT imaging datasets. Particularly, 2-D U-Net, ResU-Net++ and Dense-U-Net are implemented with the same initialization, in order to segment the LV myocardial wall and evaluate the performance of the outcomes. Both the three architectures were built in order to be functionable with CT data. Moreover, all the processing steps and the overall workflow were developed based on CT.

II. Materials and Methods

A. Dataset

The dataset used in this study, was provided by the clinical partners of SILICOFCM project. It consists of anonymized CT scans from patients with HCM. A medical expert followed a manual segmentation procedure, in order to provide the ground truth labels for the LV area upon the CT frames. The stack of the DICOM set for each patient includes 300 slices with 512x512 pixels size and a 0.4mm of slice thickness. The total number of slices is 2704 from 7 patients. To better handle and reduce the complexity, the DICOM files were converted to NIfTI files.

B. Deep Learning Network Architectures

In this study we functionally reproduce three (published in the literature) deep learning architectures of the U-Net family, in order to train them from scratch on our data. Next, a short description of the implemented networks is presented.

1) U-Net

The first U-net architecture developed by *Ronneberger et al.* [3] comprises two basic paths, the encoder and the decoder (Fig. 1 (A)). Particularly the encoder includes a convolution (and up convolution), max pooling, the ReLU as activation function at the end of each layer and a concatenation layer.

Fig. 1(A) illustrates the U-Net architecture, where the blue boxes correspond to feature maps. The number of channels is depicted on the top of each box. The white boxes are copied feature maps and the arrows denote the different operations. The convolutions are responsible to extract the feature maps and the parameters, as the pooling operations are using a filter over each feature map, in order to progressively reduce the spatial size.

2) ResU-Net++

The architecture of the ResU-Net++ is composed of the Deep Residual U-Net that exploits the strength of the deep residual learning and the U-Net. The ResU-Net++ proposed by *Jha et al.* [13] capitalizes the residual blocks, the squeeze and excitation block, Atrous Spatial Pyramidal Pooling (ASPP) and the attention block. The development of the ResU-Net++ architecture includes one stem block which is followed by 3 encoder – decoder blocks and the ASPP. The residual unit is composed of a batch normalization, ReLU activation function and convolutional layers (Fig. 1 (B)).

3) Dense-U-Net

The Dense-U-Net was developed by *Cai et al.* [14] to improve the image resolution loss from the down sampling, in order to improve the accuracy. The developed architecture consists of a combination of the U-Net model and dense concatenations. Both the dense up-sampling and dense down-sampling are symmetrical with skip connections. The Dense-U-Net is composed of 5 dense blocks both in down-sampling and up-sampling as it is depicted in Fig. 1 (C).

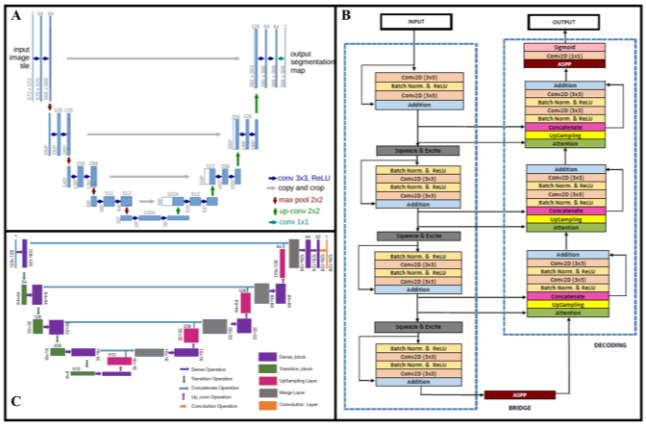

Figure 1. The proposed detailed architectures of the A) U-Net [3], B) ResU-Net++ [13] and C) Dense-U-Net [14].

C. Data preprocessing and workflow

Specific preprocessing steps are performed to enhance the CT images and visualize better the cardiac tissues and other substructures. The DICOM images and their annotations are sorted to achieve a linear match among the frames. The final output is a binary segmented mask. Fig. 2 depicts the workflow developed for the data preprocessing steps.

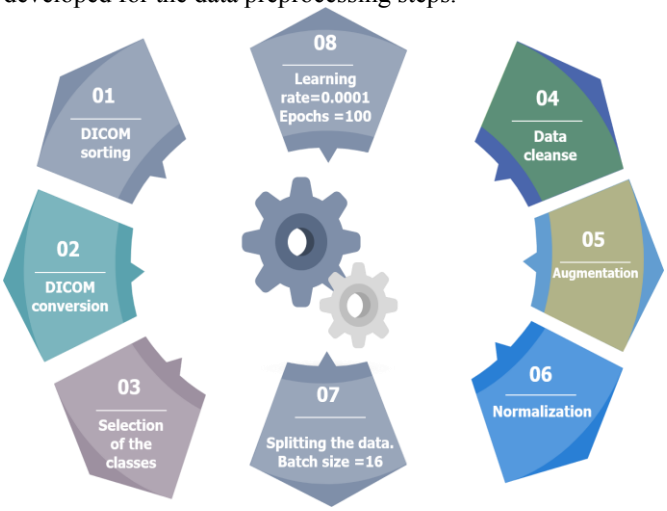

Figure 2. Workflow of the proposed pre-processing steps (1-8) before the data propagation into the network.

Below the data cleansing, augmentation, normalization and the train-test steps, are presented in detail, as they are the most important steps. Moreover, this study focus to present this steps as they were developed from scratch, in order to parameterize the approach to the targeted problem.

1) Data cleansing

First, we perform data cleansing to remove the unwanted slices (such as those with no annotations, black slices where the LV is not visible). Mostly, the DICOM images contain useful information at the middle of the stack. For that reason, mostly these images are useful. On the other hand, the images from the edges are removed when the LV is not visible and they will confuse the models. The total slices with meaningful information from the annotated masks are 1532. Also, an

important option in the data cleansing stage is the selection of the class that we want to segment or discard (LV wall or blood pool).

2) Augmentation

Prior the augmentation, a resize of the dimensions was implemented, to remove unwanted information and diminish the total extracted parameters. Afterwards, augmentation is applied, where 20% of the total slices are being shifted and rotated randomly, resulting in 1838 slices in total. The rotation is performed by randomly choosing each time at a specified angle [-20°, -10°, -5°, 5°, 10°, 20°]. The augmentation technique must be as accurate as possible in order to be helpful. Most of the times, the augmentation is a critical step as it can confuse the network instead of boosting it. For that reason, the augmentation techniques must generate high quality images.

3) Normalization

The target is to achieve a consistent intensity in the data. The normalization procedure transforms the *n*-dimensional grayscale image $I: \{X \in R^n\} \rightarrow \{Min, .., Max\}$ into a new image within the normalized range. A grayscale digital image is normalized as:

$$I_N = (I - Min)\frac{newMax - newMin}{Max - Min} + newMin, \quad (1)$$

4) Train – test split of the dataset

The learning rate for the training phase was defined 0.0001 and the batch size 16 for all networks. Regarding the LV segmentation, it is defined as a binary classification problem where we used the binary cross-entropy as a loss function defined as:

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\big(p(y_i)\big) + (1 - y_i)\log(1 - p(y_i)), \quad (2)$$

where, y is the label and $p(y)$ is the predicted probability for all N points.

To minimize any possible biased condition and evaluate the overall model performance, we performed the k-fold cross validation. The unbiased cross-validation was implemented with 5 (k) folds. The five folds were divided based on the number of frames, trying to achieve an equal spilt. This approach ensures that each fold, is an acceptable representation of the whole dataset. Successively, 5 iterations for both training and validation were performed, where at each iteration one-fold is used for the validation purposes and the rest 4 are used for the training process. The 5-fold cross validation resulted in 5 different outcomes which are summarized as mean values.

The preprocessing pipeline is an essential step, as the output improves the segmentation process. Since the data have been pre-processed, the next step is to propagate the train-test modules (CT scans and the annotations are used) through the model.

D. Comparison

The parameterization of the three utilized architectures regarding the loss function, the learning rate and the epochs is identical, in order to compare the different outcomes with the same initialization. The accuracy was the baseline metric for the outcomes and the overall performance. Moreover, the loss metric was a critical factor as it indicates the percentage of the correct predictions. Except the statistical metrics, a quantitative parameter to evaluate the image outcomes is the Dice Coefficient (DC) score. The DC score indicates the overlapping percentage of the predicted mask upon the ground truth as:

$$Dice\ coeff = \frac{2xTP}{2xTP + FP + FN}. \quad (3)$$

TP: true positive, FP=false positive and FN=false negative. The DC is the most accurate way to describe the outcome of the segmentation process.

III. Results

The testing was implemented on an Intel(R) CoreTM i7-6700HQ CPU@ 2.4Hz, using a GeForce GTX 960m GPU. The pipeline was implemented using Python 3.7, Jupyter Notebook, TensorFlow-GPU, Cuda toolkit 9.0, cuDNN v.7.0.5 and Keras. The run time takes quite big amount of time (≈ 22h) to execute, as the input size was not diminished to 256x256. The results indicated that the image preprocessing boosted and enhanced the network results. The preprocessing of the image is a critical step before the model training. High quality data with noise reduction and object removal, can improve the results of the predicted masks.

A. Evaluation metrics

Among the three used networks, the Dense-U-Net needed three times more time than the rest two, as the parameters were ten times more. Table I depicts the performance scores for all models. The Dense-U-Net architecture outperformed the other methods, as it was expected. Due to its architecture, it seeks features those other architectures do not. Dense-U-net has more complex layers where the extracted parameters from the feature maps are more.

TABLE I: METRICS FOR THE MODEL AFTER THE TRAINING PROCESS.

Networks	Metrics			
	Accuracy		*Loss*	
	Model	***Validation***	***Model***	***Validation***
U-Net	0.78	0.80	0.10	0.13
Res-U-Net++	0.80	0.80	0.15	0.17
DenseU-Net	0.81	0.82	0.23	0.2

The DC metric was calculated, in order to provide a better evaluation of the networks, as it evaluates the similarity among the predicted and the tested frames one by one. The total DC is presented with mean values, as each fold from the cross-validation outputs a separate DC value. The simple U-Net yielded a mean DC 0.81, the Res-U-Net++ 0.82 and the Dense-U-Net resulted in a DSC 0.84.

Fig. 3 depicts the predicted masks from all the implemented networks. We can observe that the masks have quite good similarity among them, but a few abnormalities appear in some frames which lead to loss. The best results were obtained when the Dense-U-Net is employed. Its architecture is more deep looking for more detailed and complicated characteristics upon the image.

The results indicated DC and performance accuracy near to 80% with data derived only from 7 patients and 2704 DICOM images. The dataset was helpful, as it consists of clean data without much noise.

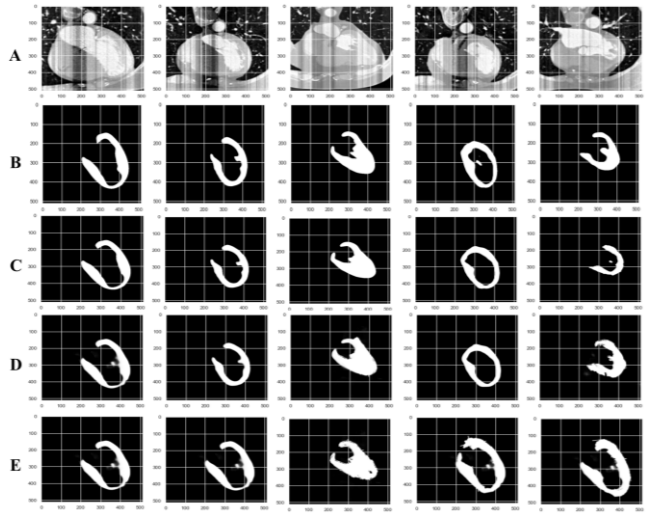

Figure 3. Results from predicted maks. A) Preprocessed CT scan, B) Ground trouth C) Dense-U-Net, E) Res-U-Net, and D) U-Net.

IV. DISCUSSION AND CONCLUSION

In summary, all the descripted networks performed very well, although the limitations of the study. With further development and refinement, they can be used for the heart segmentation task and specifically for segmenting the LV structure. The axial CT scan view was chosen in this study, as this view provides 2D masks that map the entire LV. Also, clinical metrics can be obtained (wall thickness, ejection fraction, LV volume) from this view.

In addition, the segmentation of the left ventricle is quite challenging, since there are very limited studies which focus on the axial view, as the LV shape is more complex. Using a small number of images and without optimization techniques, the results are more than acceptable. In this study, a functional comparison of the different architectures was performed. The evaluation was based on the accuracy of the predicted results and the DC. Moreover, the demonstrated DL networks are inherently good and can be extended to different applications. In conclusion, the important question to be solved in the future is how to merge large datasets with annotated labels to efficiently enhance the performance of DL networks.

For that purposes, automated segmentation can be very helpful to the clinical experts as it can provide them useful anatomical information fast and accurate.

V. LIMITATIONS

This study has few limitations starting from the data. The number of the data should be enhanced in the future in order to boost the train process. Also, the experts that provide the ground truth masks, due to lack of time, they perform a fast manual segmentation that may result to different outcomes each time. At the end, the hardware specifications must be enhanced to reduce the run time and increase the computational resources.

REFERENCES

[1] K. Kalisz and P. Rajiah, "Computed tomography of cardiomyopathies," *Cardiovasc. Diagn. Ther.*, vol. 7, no. 5, pp. 539–556, Oct. 2017, doi: 10.21037/cdt.2017.09.07.

[2] X. Zhuang and J. Shen, "Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI," *Med. Image Anal.*, vol. 31, pp. 77–87, Jul. 2016, doi: 10.1016/j.media.2016.02.006.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Lecture Notes in Computer Science*, Springer International Publishing, 2015, pp. 234–241. [Online]. Available: https://doi.org/10.1007%2F978-3-319-24574-4_28

[4] Q. Tong, M. Ning, W. Si, X. Liao, and J. Qin, "3D deeply-supervised U-net based whole heart segmentation," in *Statistical Atlases and Computational Models of the Heart: ACDC and MMWHS Challenges - 8th International Workshop, STACOM 2017, Revised Selected Papers*, Jan. 2018, pp. 224–232. doi: 10.1007/978-3-319-75541-0_24.

[5] C. Wang, T. MacGillivray, G. Macnaught, G. Yang, and D. Newby, "A two-stage 3D Unet framework for multi-class segmentation on full resolution image," *ArXiv180404341 Cs*, Apr. 2018, Accessed: Feb. 17, 2021. [Online]. Available: http://arxiv.org/abs/1804.04341

[6] Z. Xu, Z. Wu, and J. Feng, "CFUN: Combining Faster R-CNN and U-net Network for Efficient Whole Heart Segmentation," *ArXiv181204914 Cs*, Dec. 2018, Accessed: Feb. 02, 2021. [Online]. Available: http://arxiv.org/abs/1812.04914

[7] X. Zhuang *et al.*, "Evaluation of algorithms for Multi-Modality Whole Heart Segmentation: An open-access grand challenge," *Med. Image Anal.*, vol. 58, p. 101537, Dec. 2019, doi: 10.1016/j.media.2019.101537.

[8] C. Wang and O. Smedby, "Automatic Whole Heart Segmentation Using Deep Learning and Shape Context," 2018, pp. 242–249. doi: 10.1007/978-3-319-75541-0_26.

[9] A. Mortazi, J. Burt, and U. Bagci, "Multi-Planar Deep Segmentation Networks for Cardiac Substructures from MRI and CT," *ArXiv170800983 Cs Stat*, Aug. 2017, Accessed: Feb. 17, 2021. [Online]. Available: http://arxiv.org/abs/1708.00983

[10] X. Yang, C. Bian, † Lequan, D. Ni, and P.-A. Heng, *Hybrid Loss Guided Convolutional Networks for Whole Heart Parsing*. 2017.

[11] C. Ye, W. Wang, S. Zhang, and K. Wang, "Multi-Depth Fusion Network for Whole-Heart CT Image Segmentation," *IEEE Access*, vol. 7, pp. 23421–23429, 2019, doi: 10.1109/ACCESS.2019.2899635.

[12] B. Jun Guo *et al.*, "Automated left ventricular myocardium segmentation using 3D deeply supervised attention U-net for coronary computed tomography angiography; CT myocardium segmentation," *Med. Phys.*, vol. 47, no. 4, pp. 1775–1785, Apr. 2020, doi: 10.1002/mp.14066.

[13] D. Jha *et al.*, "ResUNet++: An Advanced Architecture for Medical Image Segmentation," *ArXiv191107067 Cs Eess*, Nov. 2019, Accessed: Feb. 17, 2021. [Online]. Available: http://arxiv.org/abs/1911.07067

[14] S. Cai, Y. Tian, H. Lui, H. Zeng, Y. Wu, and G. Chen, "Dense-UNet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network," *Quant. Imaging Med. Surg.*, vol. 10, no. 6, pp. 1275–1285, Jun. 2020, doi: 10.21037/qims-19-1090.