A Two-Stage Temporal-Like Fully Convolutional Network Framework for Left Ventricle Segmentation and Quantification on MR Images

Zhou Zhao, Nicolas Boutry, Élodie Puybareau, and Thierry Géraud

EPITA Research and Development Laboratory (LRDE), Le Kremlin-Bicêtre, France
elodie.puybareau@lrde.epita.fr

Abstract. Automatic segmentation of the left ventricle (LV) of a living human heart in a magnetic resonance (MR) image (2D+t) allows to measure some clinical significant indices like the regional wall thicknesses (RWT), cavity dimensions, cavity and myocardium areas, and cardiac phase. Here, we propose a novel framework made of a sequence of two fully convolutional networks (FCN). The first is a modified temporal-like VGG16 (the "localization network") and is used to localize roughly the LV (filled-in) epicardium position in each MR volume. The second FCN is a modified temporal-like VGG16 too, but devoted to segment the LV myocardium and cavity (the "segmentation network"). We evaluate the proposed method with 5-fold-cross-validation on the MICCAI 2019 LV Full Quantification Challenge dataset. For the network used to localize the epicardium, we obtain an average dice index of 0.8953 on validation set. For the segmentation network, we obtain an average dice index of 0.8664 on validation set (there, data augmentation is used). The mean absolute error (MAE) of average cavity and myocardium areas, dimensions, RWT are 114.77 mm^2; 0.9220 mm; 0.9185 mm respectively. The computation time of the pipeline is less than 2 seconds for an entire 3D volume. The error rate of phase classification is 7.6364%, which indicates that the proposed approach has a promising performance to estimate all these parameters.

Keywords: Deep learning · VGG · Left ventricle quantification · Segmentation · Fully convolutional network.

1 Introduction

Left ventricle (LV) full quantification is critical to evaluate cardiac functionality and diagnose cardiac diseases. Full quantification aims to simultaneously quantify all LV indices, including the two areas of the LV (the area of its cavity and the area of its myocardium), six RWT's (along different directions and at different positions), three LV dimensions (along different directions), and the cardiac phase (diastole or systole) [1, 2], as shown in Fig. 1. However, the LV full quantification is challenging: LV samples are variable, not only because the samples can be obtained from different hospital, but also because some of them are not concerned by cardiac diseases. It is also challenging because there are complex correlations between the LV indices. For example, the cavity area has a direct influence on the three LV dimensions and the cardiac phase.

The MICCAI 2019 Challenge on Left Ventricle Full Quantification[1] (LVQuan19) is an extension of the one of 2018 [2] with the difference that now the original data is given without preprocessing for training and testing phases, to be closer to clinical reality.

We propose then in this paper a two-stage temporal-like FCN framework that segments and estimates the parameters of interest in 2D+t sequences of the MR image of a LV. First, in each temporal frame, we localize the greatest connected component detected by the localization network, we dilate it using a size equal to 10 pixels, and we compute the corresponding bounding box. This results in a sequence of cropped LV's (that we will abusively call cropped volume). Second, we use these cropped volumes to train the LV segmentation network. The procedure is depicted in Fig. 2. Finally, the segmentation results are used for the LV full quantification.

The pipeline is based on our previous works [3, 4] but with a new step: we added one localization network before the segmentation network. Compared with [5], our localization precision is higher, because we localize the entire LV region (the filled-in epicardium) instead of the center of the bounding box containing the LV structure. Compared with [6], our method is quicker and do not have memory limit problems. To take advantages of time information, we use 3 successive 2D frames $(n-1, n, n+1)$ at time n as inputs in the localization and in the segmentation networks, yielding to better results than the traditional approach which used only the information at time n for the n^{th} slice.

We evaluated the proposed method using the dataset provided by LVQuan19 with 5-fold-cross-validation. Experiments with (very) limited training data have shown that our model has a stable performance. We added pre-processing and post-processing steps to enhance and refine our results.

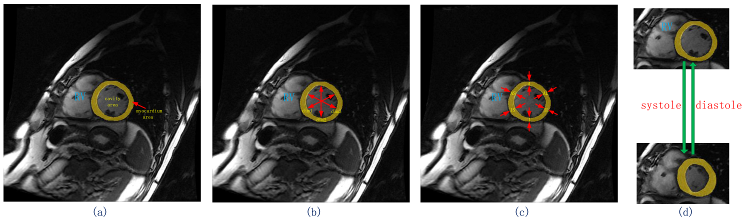


Fig. 1. Illustration of LV indices, including (a) the cavity area and the myocardium area, (b) three LV dimensions, (c) six regional wall thicknesses and (d) the cardiac phase (diastole or systole).

The plan is the following: we detail our methodology in Section 2, we detail our experiments in Section 3, and then Section 4 concludes.

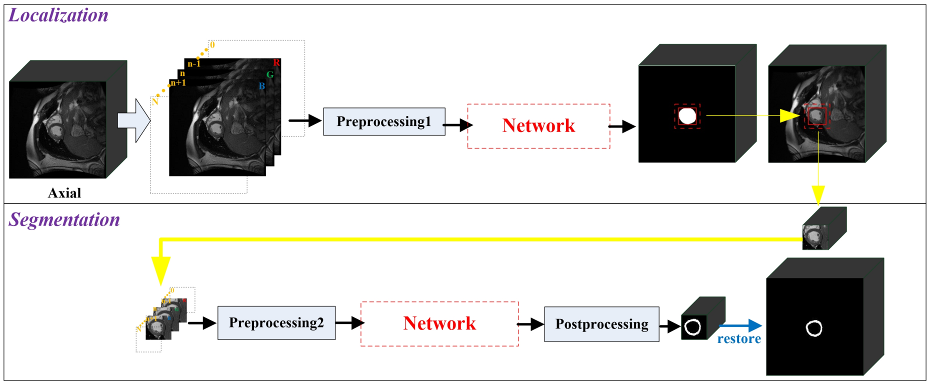


Fig. 2. Global overview of the proposed method.

2 Methodology

2.1 Dataset description

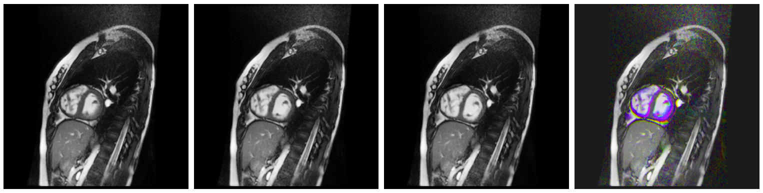
LV dataset used for this work was provided by the LVQuan19 challenge. It contains 56 patients processed SAX MR sequences. For each patient, 20 temporal frames are given and correspond to a whole cardiac cycle. All ground truth (GT) values of the LV indices are provided for every single frame. The pixel spacings of the MR images range from 0.6836 mm/pixel to 1.5625 mm/pixel, with mean values of 1.1809 mm/pixel. LV dataset includes two different image sizes: 256 × 256 or 512 × 512 pixels.

2.2 Preprocessings

Let us recall what we call *Gauss normalization*: for the $(2D+t)$-image I corresponding to a given patient, we compute $I := \frac{I-\mu}{\sigma}$ where μ is the mean of I and σ its standard deviation (σ is assumed not to be equal to zero). There are then two different pre-processing steps as depicted in Fig. 2.

- The first pre-processing (see preprocessing1 in Fig. 2) begins with a Gauss normalization. When we treat training data, we crop the initial slices into a 256 × 256 image to optimize the dice of the network (we do not do this for test datasets). Then we concatenate them for each n into a 256 × 256 × 3 pseudo-color image where R, G, B correspond respectively to $n-1, n, n+1$ (we do not detail the cases $n = 1$ and $n = 20$ because of a lack of space).
- The second pre-processing (preprocessing2 in Fig. 2) is in four steps: (1) data augmentation using rotations and flips, (2) resizing with a fixed inter-pixel spacing $(0.65mm)$, (3) Gauss normalization, and (4) we concatenate into a pseudo-color image like above.

[1] https://lvquan19.github.io
[2] https://lvquan18.github.io



(a) $n-1$ (b) n (c) $n+1$ (d) concatenation

Fig. 3. Illustration of our “temporal-like” procedure.

Because the VGG-16 network’s input is an RGB image, we propose to take advantage of the temporal information by stacking 3 successive 2D frames: to segment the n^{th} slice, we use the n^{th} slice of the MR volume, and its neighboring $(n-1)^{th}$ and $(n+1)^{th}$ slices, as green, red and blue channels, respectively. This new image, named “temporal-like” image, enhances the area of motions, here the heart, as shown in Fig. 3.

2.3 Network architecture

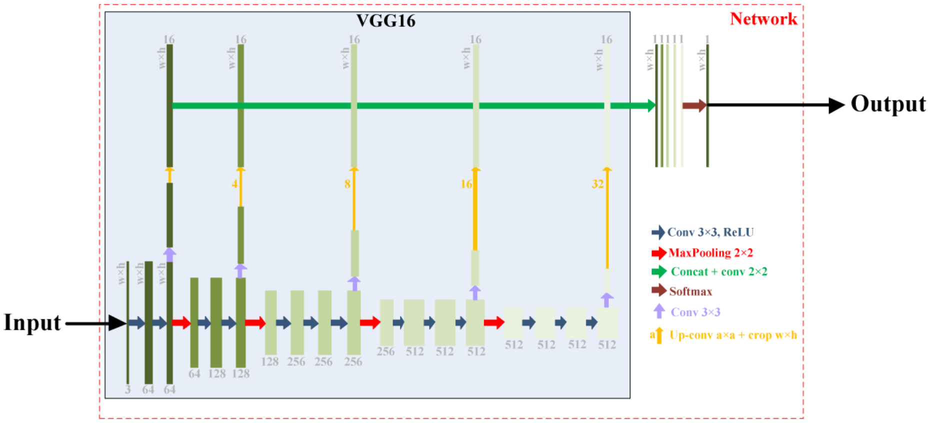


Fig. 4. Architecture of our networks.

The localization and the segmentation networks have the same architecture (see Fig. 4). First we downloaded the pre-trained original VGG16 [7] network architecture. We recall that this network has been pre-trained on millions of natural images of ImageNet for image classification [8]. Second, we discard its fully connected layers and this way we keep only the sub-network made of five convolution-based “stages” (the base network). Each stage is made of two convolutional layers, a ReLU activation function, and a max-pooling layer. Since the max-pooling layers decrease the resolution of the input image, we obtain a set of fine to coarse feature maps (with 5 levels of features). Inspired by the work in [9, 10], we added *specialized* convolutional layers (with a 3×3 kernel size) with K (*e.g.* $K = 16$) feature maps after the up-convolutional layers placed at the end of each stage. The outputs of the specialized layers have then the same resolution as the input image, and are then concatenated together. We add a 1×1 convolutional layer at the output of the concatenation layer to linearly combine the fine to coarse feature maps. This complete network provides the final segmentation result.[3]

2.4 Postprocessing

Let us assume that we input the 20 cropped temporal slices of a patient into an image of size $20 \times width \times height$ (where the crop is due to the localization procedure) in preprocessing2 to obtain a $20 \times width \times height \times 3$ image. We filter then the ouput of size $20 \times width \times height$ by keeping only the greatest connected component in the segmented $(2D + t)$-image, and we compute the inverse interpolation on the x and y axes to get back the initial inter-pixel spacing. Finally, we add a zero-valued border to get back a $20 \times 256 \times 256$ or a $20 \times 512 \times 512$ image (depending on the shape of the input).

2.5 Evaluation Methods

The LV quantification as defined in LVquan19 relies on 11 parameters: the areas of the LV cavity and the myocardium, 3 dimensions of the cavity and 6 measurements of the wall thickness. We measure the areas (see Fig. 1 (a)) by computing the number of pixels in the segmented regions corresponding to the LV cavity and the myocardium. To measure the three cavity dimension values (dim1, dim2, dim3) (see Fig. 1 (b)), we proceed this way: because our final segmentation results is the LV myocardium, we first extracted the LV cavity from the segmentation results. We then compute the boundary of the LV cavity and calculate the distances between the points of the boundary and the centroid of the LV cavity along the integral angles $\theta \in [-30, 30[$ (in degrees). Finally, we average these distances. We do this for the six separated regions of the wall. Finally, we compute the mean dimensions for each pair of opposite regions and we obtain $(dim1, dim2, dim3)$. To measure the RWT's values, we first find the boundaries of epicardium and endocardium respectively, and we compute the distances between the points on the boundary of epicardium and the points on the boundary of endocardium along the same integral angles as before where zero corresponds to the normal. Finally, we compute the mean among 60 distance values for each region. To classify the phase as systolic or diastolic, we use a simple method: we detect the time $n_{\max}$ when the cavity is maximal, and $n_{\min}$ when the cavity is minimal. Assuming that we have the case $n_{\min} > n_{\max}$, then for each time $n \in [n_{\max}, n_{\min}]$, we label the image as systolic phase, and otherwise it is a diastolic phase. We do the converse when we have $n_{\max} < n_{\min}$.

3 Experiments

We implemented our experiments on Keras/TensorFlow using a NVidia Quadro P6000 GPU. We used the multinomial logistic loss function for a one-of-many classification

[3] Note that we designed our network's architecture to work with any input shape.

task, passing real-valued predictions through a softmax to get a probability distribution over classes. For the localization network, we used an Adam optimizer (batchsize=4, β1=0.9, β2=0.999, epsilon=0.001, lr = 0.002) and we did not use learning rate decay. We trained the network during 10 epochs. We recall that we used the filled-in epicardium connected component given in the GT as the ”ones” of the output of our network. For the segmentation network, we used the same optimizer and the same parameters but we changed the batchsize to 1. Also, we considered three different classes[4] in the given GT: the background (0), the myocardium (1), the cavity (2) (we merge then 0 and 2 after the segmentation). This way, we obtained better results than using only the wall of the LV.

3.1 Results

We tested our method with 3- and 5-fold-cross-validations on the challenge dataset. An example of bounding box is depicted in red (we did not do any dilation here) in Fig. 5. We obtain an average dice index of 0.8953 on validation set. In practice, we extend next the box by a size equal to 10 pixels to ensure that the whole LV is included into the bounding box.

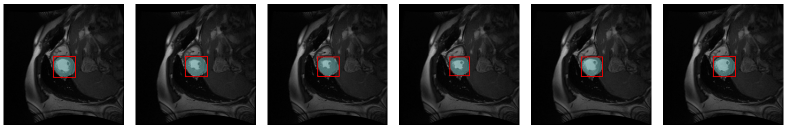


Fig. 5. Some localizations (in red) of the LV (in blue) of the 9^{th} patient.

For the segmentation, we compared ResNet50 with VGG16 as feature extraction on 3-fold-cross-validation (18, 19, 19) (see Fig. 6). VGG16 is then more efficient to detect boundaries than ResNet50 in our application.

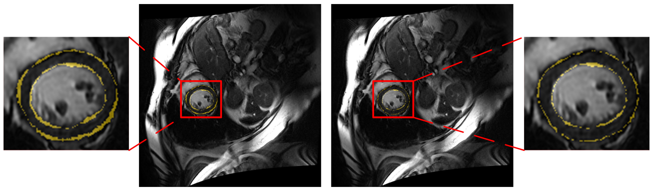


Fig. 6. Segmentation results (ResNet50-FCN on the left side vs. VGG16-FCN on the right side) for one same patient. The yellow color shows the false negatives.

Table 1 presents the average results for the two compared methods. The 11 indices of LV full quantification and dice using the VGG16-FCN are better than when we use

[4] From a technical point of view, we proceeded to a classification more than to a segmentation.

Table 1. Average results of compared methods on 3-fold-cross-validation. Values are shown as mean absolute error.

Dataset	Method	Cavity	Myocardium	Dims(mm)				RWT(mm)							Phase	Dice
		Areas(mm^2)	Areas(mm^2)	dim1	dim2	dim3	average	IS	I	IL	AL	A	AS	average	Error(%)	(%)
Validating data	ResNet50-FCN	279.32	284.84	1.8359	1.6320	1.7767	1.7482	1.2106	1.3059	1.7157	1.6225	1.3303	1.2437	1.4048	15.1267	79.20
	VGG16-FCN (our method)	88.84	157.01	0.9799	1.0691	0.9443	0.9978	0.8320	0.9173	1.1190	1.1124	0.8895	0.8408	0.9518	8.0311	86.04

the ResNet50-FCN. For these reasons, we used the VGG16-FCN for the segmentation of the LV.

To verify the stability of our algorithm, we evaluated the proposed method with 5-fold-cross-validation (11, 11, 11, 11, 12). In Table 2, the average results are showed. Compared with 3-fold-cross-validation, the average areas error is improved from 122.93 mm^2 to 114.77 mm^2, the average dims error is improved from 0.9978 mm to 0.9220 mm, the average RWT error is improved from 0.9518 mm to 0.9185 mm, the average phase error is improved from 8.0311% to 7.6364% and the dice is improved from 86.04% to 86.64%.

Table 2. Average results on 5-fold-cross-validation. Values are shown as mean absolute error.

Dataset	Cavity	Myocardium	Dims(mm)				RWT(mm)							Phase	Dice
	Areas(mm^2)	Areas(mm^2))	dim1	dim2	dim3	average	IS	I	IL	AL	A	AS	average	Error(%)	(%)
Validating data	94.31	135.23	0.9067	0.9792	0.8801	0.9220	0.8362	0.9147	1.0798	1.0560	0.8270	0.7973	0.9185	7.6364	86.64
Testing data	226.80	577.50	6.4934	3.8814	3.9835	4.7861	4.2693	1.8585	2.0570	1.9129	1.6441	3.6039	2.5576	9.83	-

In Table 2, we also reported the results on test dataset given by the organizers of LVQuan19. The test dataset was composed of processed SAX MR sequences of 30 patients. For each patient, only the SAX image sequences of 20 frames were provided (no GT).

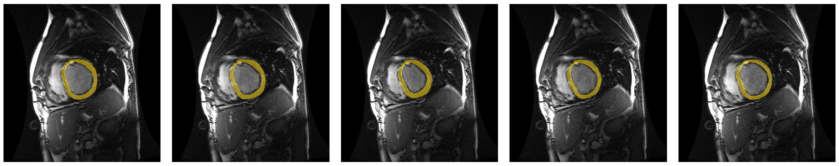


Fig. 7. Some segmentation results on the 5^{th} patient of test dataset.

In Fig. 7, the segmentation results on fifth patient of test dataset are showed, the yellow ring denotes the segmentation results.

4 Conclusion

In this paper, we propose to use a modified VGG16 to proceed to pixelwise image segmentation, in particular to segment the wall of the heart LV in temporal MR images. The proposed method provides promising results at the same time in matter of localization and segmentation, and leads to realistic physical measures of clinical values relative to the human heart. Our perspective is to try to better segment the boundary of the wall of the LV, either by increasing the weights relative to the boundary regions in the loss function, or by separating the boundary and the interior of the wall into two classes during the classification procedure.

Acknowledgements We thank the organizers of the MICCAI 2019 LV Full Quantification Challenge for providing the LV dataset, NVidia for giving us a Quadro P6000 GPU for this research, and the financial support from China Scholarship Council (CSC, File No.201806290010)

References

1. Xue, W. F., Brahm, G., Pandey, S., Leung, S., Li, S.: Full left ventricle quantification via deep multitask relationships learning. Med. Image Anal. **43**, 54–65 (2018).
2. Xue, W. F., Lum, A., Mercado, A., Landis, M., Warringto, J., Li, S.: Full quantification of left ventricle via deep multitask learning network respecting intra-and inter-task relatedness. In: Descoteaux, M.Maier-Hein, L., Franz, A., Jannin, P., Collins, P.,Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 276–284. Springer, Cham (2017). https://doi.org/10.1007/978–3–319–66179–7_32
3. Xu, Y., Géraud, T., Bloch, I.: From neonatal to adult brain MR image segmentation in a few seconds using 3D-like fully convolutional network and transfer learning, Proc. of ICIP, pp.4417–4421. IEEE, Beijing (2017). https://doi.org/10.1109/$ICIP$.2017.8297117
4. Puybareau, E., Zhao, Z., Khoudli, Y., Carlinet, E., Xu Y. C., Lacotte J., Géraud T.: Left atrial segmentation in a few seconds using fully convolutional network and transfer learning, In: Pop, M., Sermesant M.,Zhao J. C., Li, S., McLeod, K., Young, A., Rhode, K., Mansi, T. (eds.) STACOM 2018. LNCS, vol. 11395, pp. 339–347. Springer, Cham (2018). https://doi.org/10.1007/978–3–030–12029–0_37
5. Payer, C.,Stern, D., Bischof, H., Carlinet, E., Urschler, M.: Multi-label Whole Heart Segmentation Using CNNs and Anatomical Label Configurations, In: Pop M., Sermesant, M., Jodoin, P. M., Lalande, A., Zhuang, X. H., Yang, G., Young, A., Bernard, O.(eds.) STACOM 2017. LNCS, vol. 10663, pp. 190–198. Springer, Cham (2017). https://doi.org/10.1007/978–3–319–75541–0_20
6. Wang, C. J., MacGillivray, T., Macnaught, G., Yang, G., Newby, D.: A two-stage 3D Unet framework for multi-class segmentation on full resolution image. CoRR abs/1804.04341 (2018)
7. Simonyan, K., Zisserman A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014)
8. Krizhevsky, A., Sutskever, I., Hinton G. E.: ImageNet classification with deep convolutional neural networks. Advances in neural information processing systems, pp. 1097–1105, 2012
9. Long J., Shelhamer E., Darrell T.: Fully convolutional networks for semantic segmentation. Proc. of CVPR, pp.3431–3440. IEEE, Boston (2015).
10. Maninis, K.K.,Pont-Tuset, J.,Arbeláez, P., Van Gool, L.: Deep Retinal Image Understanding. In: Ourselin, S., Joskowicz, L., Sabuncu, M., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9351, pp. 140–148. Springer, Cham (2016). https://doi.org/10.1007/978–3–319–46723–8_17