Left ventricle quantification through spatio-temporal CNNs

Alejandro Debus, Enzo Ferrante

Research institute for signals, systems and computational intelligence, *sinc(i)*,
FICH-UNL/CONICET
(Santa Fe, Argentina)

Abstract. Cardiovascular diseases are among the leading causes of death globally. Cardiac left ventricle (LV) quantification is known to be one of the most important tasks for the identification and diagnosis of such pathologies. In this paper, we propose a deep learning method that incorporates 3D spatio-temporal convolutions to perform direct left ventricle quantification from cardiac MR sequences. Instead of analysing slices independently, we process stacks of temporally adjacent slices by means of 3D convolutional kernels which fuse the spatio-temporal information, incorporating the temporal dynamics of the heart to the learned model. We show that incorporating such information by means of spatio-temporal convolutions into standard LV quantification architectures improves the accuracy of the predictions when compared with single-slice models, achieving competitive results for all cardiac indices and significantly breaking the state of the art [10] for cardiac phase estimation.

Keywords: Left ventricle quantification, Spatio temporal convolutional neural network

1 Introduction

In 2015, around 17.7 million people died worldwide due to heart diseases. Left ventricle (LV) quantification is a key factor for the identification and diagnosis of such pathologies [2]. However, the estimation of cardiac indices remains a very complex task due to its intricated temporal dynamics and the inter-subject variability of the cardiac structures. Indices such as cavity and myocardium area, regional wall thickness, cavity dimensions, among others, provide useful information to diagnose various types of cardiac pathologies. Cardiovascular magnetic resonance (CMR) is one of the preferred modalities for LV related studies since it is non invasive, presents high spatio-temporal resolution, has a good signal-to-noise ratio and allows to clearly identify the tissues and muscles of interest [6].

The classical approach to LV quantification consists in estimating such indices by means of automatic segmentation [3–7, 9]. Segmentation is usually performed following supervised learning approaches, which require expert manual annotations contouring the edges of the myocardium for training. Once the segmentation is performed, the indices are computed from the resulting mask. Therefore,

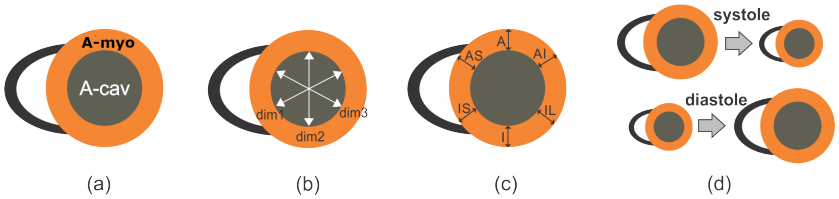

Fig. 1: Illustration of indices of the left cardiac ventricle (based on Fig. 1 from [10]). (a) Cavity area (brown) and myocardial area (orange). (b) Directional dimensions of cavity (white arrows). (c) Regional wall thicknesses. A: anterior; AS: anterospetal; IS: inferoseptal; I: inferior; IL: inferolateral; AL: anterolateral. (d) Cardiac phase (systole or diastole)

the accuracy of the predicted indices is conditioned on the quality of the segmentation. In this work, we follow an alternative strategy that directly estimates the indices of interest from the input image sequence. Inspired by the work of [11, 10, 12], our model is based on a convolutional neural network directly operating on images and regressing the target indices. Different from previous approaches like [10] where the temporal dynamics of cardiac sequences is incorporated using recurrent neural networks (RNNs), we propose a simple but effective strategy based on the use of spatio-temporal convolutions [8]. In the context of video analysis, spatio-temporal convolutions are standard 3D convolutions that operate on spatio-temporal video volumes [7]. Here we employ them to process subsets of temporally contiguous CMR slices, leveraging temporal information towards improving prediction accuracy.

We investigate the use of spatio-temporal convolutions for estimating cardiac phase, directional dimensions of the cavity, regional wall thicknesses and area of cavity and myocardium under the hypothesis that such indices may be better explained when taking into account the temporal dynamics of the heart. We benchmark the proposed architecture using the LVQuan Challenge 2018[1] dataset, which provides CMR sequences with annotations for the aforementioned indices, and provide empirical evidence that incorporating the temporal dynamics of the heart through 3D spatio-temporal convolutions improves prediction accuracy when compared with single-slice models.

2 Materials and methods

2.1 Architecture

An overview of the proposed CNN architecture is presented in Figure 2. The network takes sequences of κ slices and outputs the corresponding indices *only* for the central slice. In such way, we incorporate information from the surrounding slices, easing the prediction task. In what follows, we describe in detail the

[1] LVQuan Challenge website: https://lvquan18.github.io/

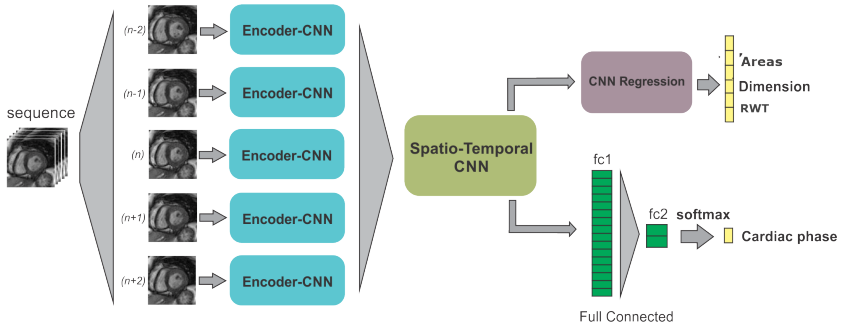

Fig. 2: Overview of proposed architecture.

main components of the proposed architecture.

Encoder-CNN. We use a first CNN (referred as encoder-CNN in Figures 2 and 3) to extract informative features from individual slices. Inspired by [11], we designed the per-slice encoding phase using a two-layers CNN where the convolutional and pooling kernels are of size 5x5, instead of the frequently used 3x3, to introduce more shift invariance (see Figure 3 for more details). We use ReLU activation function and batch normalization to alleviate the training process.

Spatio-Temporal CNN. After the encoding phase, the 40 filters generated for every individual encoder-CNN are used to construct a spatio-temporal volume with 40 channels per temporal slice. This volume is then processed using 3D convolutions that operate on the temporal and spatial dimensions (see Figure 3), producing compound feature maps that incorporate information from both of them. This module is composed of two 3D convolutional layers with kernels of size 3x5x5 and 2x5x5 when considering $\kappa = 5$ slices. When considering $\kappa = 1, 3, 7$ slices, the proposed architecture is modified by using padding in the temporal dimension ($\kappa = 1, 3$) and adding an extra convolution ($\kappa = 7$) so that the shape of the output tensor matches 1x6x6, the size required by the CNN Regression and Fully Connected modules. ReLU activations and batch normalization are also used in this module.

Final parallel branches. After fusing the spatio-temporal features, two parallel branches are derived: (i) the first branch corresponds to a shallow CNN coupled after the spatio-temporal module, acting as a regressor of the directional dimensions, wall thickness and areas; (ii) in the second branch, a third convolutional layer is coupled to the spatio-temporal module, followed by a fully connected multi layer perceptron (MLP) with 640 neurons in the hidden layer and 2 output neurons encoding the probability for the cardiac phase (systole or diastole).

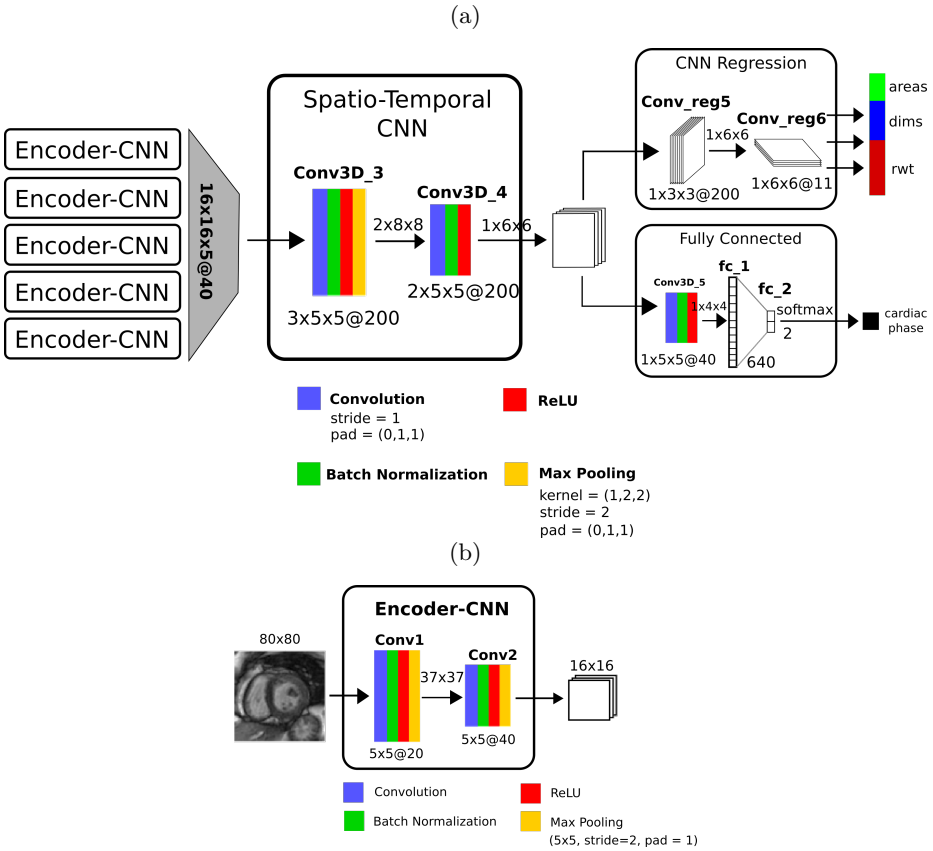

Fig. 3: (a) Deatiled overview of the spatio-temporal CNN based on 3D convolutions. (b) Zoomed version of the individual encoder-CNNs: for a single input slice of size 80x80 it outputs 40 filters of size 16x16 which are then fed to the spatio-temporal CNN.

Training procedure and loss function. We train the proposed network by minimizing a loss function over sets of κ slices where annotations are provided only for the central slice. Given a set of κ slices $\boldsymbol{x^i} = \{x_0, ..., x_\kappa - 1\}$, ground-truth annotations for the central slice $\boldsymbol{y^i} = \{y_{dim}, y_{areas}, y_{rwt}, y_{phase}\}$ and corresponding predictions from the proposed neural network ϕ_{phase} and $\phi_{dim}, \phi_{areas}, \phi_{rwt}$ the loss function is defined as:

$$\mathcal{L}(\boldsymbol{x^i}, \boldsymbol{y^i}) = \mathcal{L}_{mse}(\phi_{areas}, y_{areas}) + \mathcal{L}_{mse}(\phi_{dim}, y_{dim}) + \\ \mathcal{L}_{mse}(\phi_{rwt}, y_{rwt}) + \mathcal{L}_{ce}(\phi_{phase}, y_{phase}) + \lambda \mathcal{L}_{reg}, \tag{1}$$

where $\mathcal{L}_{mse}$ is the mean squared error between predictions and ground truth, $\mathcal{L}_{ce}$ is the cross-entropy loss, $\mathcal{L}_{reg}$ is the regularizer (L2 norm of the network

weights) and λ is a weighting factor. We minimize this loss using stochastic gradient descent with momentum, with mini-batches of size $s = 20$.

Circular hypothesis. Since we require sets of temporally contiguous slices as input for our spatio-temporal architecture, given a sequence of N slices, we adopt a circular hypothesis meaning that slice number $N - 1$ is temporally followed by slice 0. This hypothesis was corroborated by visual inspection of the training dataset. Following this strategy, we generate sets of κ slices for every sequence and use them as independent data samples. At prediction time, we employ the same hypothesis to generate the sets of test slices.

2.2 Dataset and experimental setting

Our method is experimentally validated using the training data provided by the LVQuan challenge 2018, composed of short axis cardiac MR images of 145 subjects. For each subject, it contains 20 frames corresponding to a complete cardiac cycle (giving a total of 2900 images in the dataset with pixel spacing ranging from 0.6836 mm/pixel to 2.0833 mm/pixel, with a mean of 1.5625 mm/pixel). The images have been collected from 3 different hospitals and subjects are between 16 and 97 years of age, with an average of 58.9 years. All cardiac images undergo several preprocessing steps (including historical tagging, rotation, ROI clipping, and resizing). The resulting images are roughly aligned with a dimension of 80x80. Epicardium and endocardium borders were manually annotated by radiologists, and used to extract the ground truth LV indices and cardiac phase. The values of regional wall thickness and the dimensions of the cavity are normalized by the dimension of the image, while the areas are normalized by the pixel number (6400).

In our experiments, we used cross validation with 3, 5 and 7 folds as suggested by the LVQuan organizers, resulting in partitions of size (49, 48, 48), (29, 29, 29, 29, 29) and (21, 21, 21, 21, 21, 20, 20) respectively. We used learning rate = 1e-4, momentum = 0.5 and $\lambda = 0.005$ (these parameters were obtained by grid-search).

The model was implemented in Python[2], using PyTorch and trained in GPU.

Evaluation criteria. Pearson correlation coefficient (PCC) and Mean Absolute Error (MAE) were used to assess the performance of the algorithms for estimation of areas, dimensions and regional wall thicknesses. Error Rate (ER) was used to assess the performance for cardiac phase classification.

$$PCC_{ind} = \frac{\sum_{i=1}^{N}(\phi_{ind}^{(i)} - \bar{\phi}_{ind})(y_{ind}^{(i)} - \bar{y}_{ind})}{\sqrt{\sum_{i=1}^{N}(\phi_{ind}^{(i)} - \bar{\phi}_{ind})^2 \sum_{i=1}^{N}(y_{ind}^{(i)} - \bar{y}_{ind})^2}}, \tag{2}$$

[2] The source code for the proposed architecture is publicly available at https://github.com/alejandrodebus/SpatioTemporalCNN_lvquan

$$MAE_{ind} = \frac{1}{N}\sum_{i=1}^{N}|\phi_{ind}^{(i)} - y_{ind}^{(i)}|, \tag{3}$$

where $ind \in (A_1, A_2, D_1...D_3, RWT_1...RWT_6)$, y_{ind} is the ground-truth value and ϕ_{ind} is the estimated value.$\bar{y}_{ind}$ and $\bar{\phi}_{ind}$ are their mean values, respectively.

$$ER_{phase} = \frac{\sum_{i=1}^{N}\mathbf{1}(\phi_{phase}^{(i)} \neq y_{phase}^{(i)})}{N}100\% \tag{4}$$

where $\mathbf{1}()$ is the indication function, ϕ_{phase} and y_{phase} are the estimated and ground truth value of the cardiac phase, respectively.

3 Results and discussion

The effectiveness of the proposed method was validated under the experimental setting discussed in Section 2.2. We measured the influence of the parameter κ (number of contiguous slices fed to the network) for $\kappa = 1$ (single slice), 3, 5 and 7 for the proposed spatio-temporal model based on 3D convolutions, and compare with the state of the art method recently proposed in [10]. Results are presented in Table 1 for a 5-fold cross validation setting (the same experimental setting and dataset was used in [10]). Note that using sets of $\kappa = 5$ slices significantly outperforms the configurations $\kappa = 1, 3$ for all the indices, highlighting the importance of the temporal dynamics. However, considering $\kappa = 5$ and $\kappa = 7$ slices achieves a similar performance. Therefore, we consider $\kappa = 5$ as enough temporal context for the remaining experiments.

In quantitative terms, we reduce the error rate from 28.45.06% to 3.85% for cardiac phase estimation and the MAE from 270 to $190mm^2$, 3.18 to $2.29mm$ and 2.62 to $1.42mm$ in average for the areas, directional dimensions of the cavity and regional wall thickness when comparing the performance for $\kappa = 1$ and $\kappa = 5$ slices respectively. Moreover, considering the baseline [10] we observe similar results for most indices, expect for the phase, where our model improves over the state of the art by a significant margin (reducing the error rate from 8.2% to 3.2%)

Finally, table 2 presents these results for 3 different cross-validation configurations (3, 5 and 7 folds) as required by the LVQuan challenge organizers, together with the results for phase, directional dimensions, regional wall thicknesses and area of cavity and myocardium obtained with the best performing spatio-temporal model ($\kappa = 5$). Note that performance is consistent across folds.

	Model				
	κ=1	κ=3	$\kappa = \mathbf{5}$	κ=7	DMTRL [10]
	Areas (mm^2)				
a-cav	239 ± 198	194 ± 188	181 ± 130	180 ± 145	$\mathbf{172 \pm 148}$
	0.861 ± 0.053	0.922 ± 0.035	0.940 ± 0.014	0.923 ± 0.016	$\mathbf{0.943}$
a-myo	301 ± 243	223 ± 179	199 ± 138	207 ± 141	$\mathbf{189 \pm 159}$
	0.852 ± 0.047	0.892 ± 0.029	0.923 ± 0.016	0.931 ± 0.017	$\mathbf{0.947}$
average	270 ± 154	208 ± 141	190 ± 122	193 ± 115	$\mathbf{180 \pm 118}$
	0.857 ± 0.049	0.907 ± 0.033	0.932 ± 0.015	0.927 ± 0.018	$\mathbf{0.945}$
	Dimensions (mm)				
dim1	3.05 ± 2.84	2.63 ± 2.01	$\mathbf{2.27 \pm 1.79}$	2.31 ± 1.81	2.47 ± 1.95
	0.861 ± 0.031	0.925 ± 0.018	$\mathbf{0.961 \pm 0.012}$	0.952 ± 0.015	0.957
dim2	3.23 ± 3.02	2.80 ± 1.89	$\mathbf{2.38 \pm 1.90}$	2.41 ± 2.03	2.59 ± 2.07
	0.879 ± 0.061	0.932 ± 0.023	0.957 ± 0.012	$\mathbf{0.961 \pm 0.013}$	0.894
dim3	3.27 ± 3.12	2.56 ± 1.75	$\mathbf{2.22 \pm 1.78}$	2.23 ± 1.67	2.48 ± 2.34
	0.912 ± 0.047	0.939 ± 0.021	$\mathbf{0.963 \pm 0.011}$	0.959 ± 0.010	0.943
average	3.18 ± 2.54	2.66 ± 1.75	$\mathbf{2.29 \pm 1.59}$	2.31 ± 1.62	2.51 ± 1.58
	0.884 ± 0.044	0.932 ± 0.022	$\mathbf{0.960 \pm 0.011}$	0.957 ± 0.012	0.925
	Regional wall Thickness (mm)				
wt1 (IS)	2.02 ± 1.32	1.89 ± 1.04	$\mathbf{1.23 \pm 1.14}$	1.24 ± 1.17	1.26 ± 1.04
	0.625 ± 0.063	0.793 ± 0.056	0.854 ± 0.014	0.846 ± 0.011	$\mathbf{0.856}$
wt2 (I)	2.67 ± 1.69	2.45 ± 1.48	1.44 ± 1.22	1.43 ± 1.87	$\mathbf{1.40 \pm 1.10}$
	0.618 ± 0.055	0.751 ± 0.037	0.797 ± 0.011	$\mathbf{0.801 \pm 0.014}$	0.747
wt3 (IL)	2.95 ± 2.01	1.74 ± 1.56	$\mathbf{1.57 \pm 1.41}$	1.60 ± 1.59	1.59 ± 1.29
	0.595 ± 0.049	0.735 ± 0.033	$\mathbf{0.765 \pm 0.013}$	0.740 ± 0.010	0.693
wt4 (AL)	2.77 ± 1.65	1.66 ± 1.17	1.48 ± 1.13	$\mathbf{1.46 \pm 1.45}$	1.57 ± 1.34
	0.603 ± 0.052	0.763 ± 0.024	$\mathbf{0.785 \pm 0.022}$	0.782 ± 0.018	0.659
wt5 (A)	3.06 ± 2.12	1.49 ± 1.35	1.35 ± 1.19	1.39 ± 1.21	$\mathbf{1.32 \pm 1.10}$
	0.642 ± 0.061	0.808 ± 0.029	0.842 ± 0.019	$\mathbf{0.851 \pm 0.015}$	0.777
wt6 (AS)	2.25 ± 1.72	1.65 ± 1.11	1.46 ± 1.32	1.49 ± 1.37	$\mathbf{1.25 \pm 1.01}$
	0.651 ± 0.047	0.825 ± 0.032	0.870 ± 0.015	0.866 ± 0.013	$\mathbf{0.877}$
average	2.62 ± 2.10	1.81 ± 1.05	1.42 ± 0.65	1.43 ± 0.71	$\mathbf{1.39 \pm 0.68}$
	0.622 ± 0.054	0.779 ± 0.033	$\mathbf{0.819 \pm 0.015}$	0.814 ± 0.014	0.768
	Phase (ER%)				
phase	28.45 ± 5.50	14.67 ± 3.65	$\mathbf{3.85 \pm 2.82}$	3.91 ± 2.76	8.2

Table 1: Sensitivity analysis for the parameter κ (number of neighbouring slices) when using the spatio-temporal model based on 3D convolutions with 5-folds cross validation, compared with the state of the art DMTRL proposed in [10]. Note that incorporating the temporal dynamics by considering multiple slices ($\kappa = 3, 5, 7$) makes a significant different with respect the single slice case ($\kappa = 1$). However, considering $\kappa = 5$ and $\kappa = 7$ slices present a similar performance. Therefore, we consider $\kappa = 5$ as enough temporal context for the remaining experiments. When comparing with [10] we observe similar results for most indices, expept for the phase, where the proposed model breaks the state of the art significantly (from 8.2% to 3.2%).

4 Conclusions

In this work, we proposed a new CNN architecture for LV quantification that incorporates the dynamics of the heart by means of spatio-temporal convolutions. Differently from other methods that rely on more complex mechanisms (like recurrent neural networks [10]) we employ simple 3D convolutions to fuse information coming from temporally contiguous CMR slices. We generated training samples following a circular hypothesis, meaning that first and last slices of the sequences are considered as temporally contiguous. Validation was performed using CRM sequences provided by the LVQuan challenge organizers. Results show that incorporating temporal information through spatio-temporal convolutions significantly boosts prediction performance for all the indices. Moreover, when compared with the RNN based model presented in [10], we observe a significant reduction in error rate for phase estimation (from 8.2% to 3.85%) while keeping equivalent results for the other indices. More importantly, our method achieves state of the art results employing simple 3D convolutions instead of the more complex parallel RNN and Bayesian based multitask relationship learning module proposed in [10].

In this work we incorporated the spatio-temporal dynamics by means of 3D convolutions. However, if we consider the slices as multiple channels of a standard 2D architecture, conventional 2D convolutions could also be used, reducing the complexity of the model. Moreover, temporal information encoded by inter-slice deformation fields (obtained trough deep learning based image registration methods [1]) could also be considered to improve model performance. In the future, we plan to explore the performance of these models when compared with the proposed architecture.

Acknowledgements

The present work used computational resources of the Pirayu Cluster, acquired with funds from the Santa Fe Science, Technology and Innovation Agency (ASACTEI), Government of the Province of Santa Fe, through Project AC-00010-18, Resolution N 117/14. This equipment is part of the National System of High Performance Computing of the Ministry of Science, Technology and Productive Innovation of the Republic of Argentina. We also thank NVidia for the donation of a GPU used for this project. Enzo Ferrante is a beneficiary of an AXA Research Fund grant.

References

1. Ferrante, E., Oktay, O., Glocker, B., Milone, D.: On the adaptability of unsupervised cnn-based deformable image registration to unseen image domains. In: 9th International Conference on Machine Learning in Medical Imaging (MLMI 2018 - MICCAI) (2018 - In Press)

N-fold cross validation as required by LVQuan Challenge						
	MAE			PCC		
	N=3	N=5	N=7	N=3	N=5	N=7
Areas (mm^2)						
a-cav	185 ± 125	181 ± 130	183 ± 115	0.932	0.940	0.939
a-myo	204 ± 143	199 ± 138	198 ± 145	0.915	0.923	0.930
average	194 ± 131	190 ± 122	190 ± 110	0.924	0.932	0.935
Dimensions (mm)						
dim1	2.71 ± 2.11	2.27 ± 1.79	2.26 ± 1.82	0.938	0.961	0.959
dim2	2.65 ± 2.09	2.38 ± 1.90	2.32 ± 2.01	0.926	0.957	0.954
dim3	2.51 ± 2.20	2.22 ± 1.78	2.24 ± 1.91	0.933	0.963	0.958
average	2.62 ± 1.87	2.29 ± 1.59	2.27 ± 1.52	0.932	0.960	0.957
Regional wall Thickness (mm)						
wt1 (IS)	1.31 ± 1.16	1.23 ± 1.14	1.25 ± 1.15	0.831	0.854	0.857
wt2 (I)	1.58 ± 1.10	1.44 ± 1.22	1.43 ± 1.41	0.768	0.797	0.802
wt3 (IL)	1.62 ± 1.22	1.57 ± 1.41	1.56 ± 1.56	0.743	0.765	0.755
wt4 (AL)	1.60 ± 1.08	1.48 ± 1.13	1.50 ± 1.11	0.776	0.785	0.797
wt5 (A)	1.43 ± 1.12	1.35 ± 1.19	1.33 ± 1.24	0.829	0.842	0.861
wt6 (AS)	1.52 ± 1.29	1.46 ± 1.32	1.46 ± 1.09	0.857	0.870	0.873
average	1.51 ± 0.98	1.42 ± 0.65	1.42 ± 0.61	0.801	0.819	0.824
Phase (ER %)						
	N=3		N=5		N=7	
phase	5.10 ± 3.72		3.85 ± 2.82		3.81 ± 3.05	

Table 2: Results obtained for the LVQuan challenge dataset using the proposed spatio-temporal model (areas of LV cavity and myocardium (mm^2), directional dimensions (mm), wall thicknesses (mm) and cardiac phase) for N-folds cross validation with $N = 3, 5$ and 7.

2. Karamitsos, T.D., Francis, J.M., Myerson, S., Selvanayagam, J.B., Neubauer, S.: The role of cardiovascular magnetic resonance imaging in heart failure. Journal of the American College of Cardiology 54(15), 1407–1424 (2009)
3. Peng, P., Lekadir, K., Gooya, A., Shao, L., Petersen, S.E., Frangi, A.F.: A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging. Magnetic Resonance Materials in Physics, Biology and Medicine 29(2), 155–195 (2016)
4. Petitjean, C., Dacher, J.N.: A review of segmentation methods in short axis cardiac mr images. Medical image analysis 15(2), 169–184 (2011)
5. Poudel, R.P., Lamata, P., Montana, G.: Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation. In: Reconstruction, Segmentation, and Analysis of Medical Images, pp. 83–94. Springer (2016)
6. Suinesiaputra, A., Bluemke, D.A., Cowan, B.R., Friedrich, M.G., Kramer, C.M., Kwong, R., Plein, S., Schulz-Menger, J., Westenberg, J.J., Young, A.A., et al.: Quantification of lv function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours. Journal of Cardiovascular Magnetic Resonance 17(1), 63 (2015)

7. Tan, L.K., Liew, Y.M., Lim, E., McLaughlin, R.A.: Convolutional neural network regression for short-axis left ventricle segmentation in cardiac cine mr sequences. Medical image analysis 39, 78–86 (2017)
8. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Computer Vision (ICCV), 2015 IEEE International Conference on. pp. 4489–4497. IEEE (2015)
9. Tran, P.V.: A fully convolutional neural network for cardiac segmentation in short-axis mri. arXiv preprint arXiv:1604.00494 (2016)
10. Xue, W., Brahm, G., Pandey, S., Leung, S., Li, S.: Full left ventricle quantification via deep multitask relationships learning. Medical image analysis 43, 54–65 (2018)
11. Xue, W., Lum, A., Mercado, A., Landis, M., Warrington, J., Li, S.: Full quantification of left ventricle via deep multitask learning network respecting intra-and inter-task relatedness. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 276–284. Springer (2017)
12. Xue, W., Nachum, I.B., Pandey, S., Warrington, J., Leung, S., Li, S.: Direct estimation of regional wall thicknesses via residual recurrent neural network. In: International Conference on Information Processing in Medical Imaging. pp. 505–516. Springer (2017)