Left Ventricle Quantification with Cardiac MRI: Deep Learning Meets Statistical Models of Deformation

Jorge Corral Acero[1(✉)], Hao Xu[1], Ernesto Zacur[1], Jurgen E. Schneider[2], Pablo Lamata[3], Alfonso Bueno-Orovio[4], and Vicente Grau[1]

[1] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK
{jor.corral,hao.xu}@eng.ox.ac.uk
[2] Leeds Institute of Cardiovascular and Metabolic Medicine, University of Leeds, Leeds, UK
[3] Department of Biomedical Engineering, King's College of London, London, UK
[4] Department of Computer Science, University of Oxford, Oxford, UK

Abstract. Deep learning has been widely applied for left ventricle (LV) analysis, obtaining state of the art results in quantification through image segmentation. When the training datasets are limited, data augmentation becomes critical, but standard augmentation methods do not usually incorporate the natural variation of anatomy. In this paper we propose a pipeline for LV quantification applying our data augmentation methodology based on statistical models of deformations (SMOD) to quantify LV based on segmentation of cardiac MR (CMR) images, and present an in-depth analysis of the effects of deformation parameters in SMOD performance. We trained and evaluated our pipeline on the MICCAI 2019 Left Ventricle Full Quantification Challenge dataset, and achieved average mean absolute error (MAE) for areas, dimensions, regional wall thickness and phase of 106 mm^2, 1.52 mm, 1.01 mm and 8.0% respectively in a 3-fold cross-validation experiment.

Keywords: Deep learning · Data augmentation · LV quantification

1 Introduction

Automatic quantification of the left ventricle (LV) has been greatly enhanced by the development of deep learning algorithms in the past few years. Convolutional neural networks have shown great accuracy and flexibility for LV quantification. Recently, the MICCAI 2018 Left Ventricle Full Quantification Challenge made possible to compare a wide range of deep learning algorithms performing on the same benchmark dataset with both direct regression [1] and segmentation based [2–4] approaches. Direct

Jorge Corral Acero and Hao Xu contributed equally.

M. Pop et al. (Eds.): STACOM 2019, LNCS 12009, pp. 384–394, 2020.
https://doi.org/10.1007/978-3-030-39074-7_40

regression approaches have shown promising results, while segmentation-based approaches were in general, at the time of the challenge, more accurate.

With the development of big databases such as UK Biobank [5], applying deep learning algorithms on big data has become possible in biomedical applications [6, 7], influencing the choice and design of neural networks. With more training data, deeper networks with more parameters can be trained, which usually results in better performance. However, in clinical practice, especially for pathological cases, it is difficult to acquire such big dataset, and data augmentation becomes important. In this regard, our recently developed augmentation method based on statistical models of deformation has shown promising results on a variety of datasets for segmentation task [8].

The MICCAI 2019 Left Ventricle Full Quantification Challenge has provided a benchmark dataset which, compared to the corresponding 2018 dataset, is closer to real-life clinical conditions, with no pre-processing applied to the images. We propose a segmentation-based quantification pipeline enhanced with statistical models of deformation, developed and evaluated on this dataset.

2 Methods

We propose a complete pipeline for quantifying the LV from cardiac MR (CMR) images, consisting of the following steps. We first build a population-specific atlas, and train an initial neural network to locate the centre of the heart in all the images. We then rigidly register each image to the atlas previously calculated. We build the statistical models of deformation, which we use to augment the images using different strategies. Finally, we train a second neural network to perform the fine segmentation and retrieve the LV metrics from the segmentation results.

2.1 Data

We developed and evaluated our pipeline using the MICCAI 2019 Left Ventricle Full Quantification Challenge dataset, which consists of 56 training subjects and 30 testing subjects. For each subject in the training data, a single short-axis (SAX) CMR sequence consisting of 20 frames was provided together with a set of clinically significant LV indices including regional wall thicknesses, cavity dimensions, cavity areas and myocardium and cardiac phase for each frame. Endocardial and epicardial segmentation binary masks were also made available as reference, and pixel-spacing values were also given for metrics evaluation. For subjects in the testing dataset, only CMR image sequences and pixel-spacing values were provided.

Comparing to MICCAI 2018 Left Ventricle Full Quantification Challenge, which had 145 training subjects and 30 testing subjects, the size of training dataset reduced by 61.4% and the testing dataset remained the same size [10].

2.2 Rigid Registration

Our rigid registration method was based on the maximization of cross-correlation of image intensities. In order to avoid converging to a local minimum, the algorithm was

initialized to different transformations distributed in the space of possible transformations. Diffeomorphic Log Demons [11] was applied for non-rigid registration ($\sigma_{fluid} = 2, \sigma_{diff} = 1.8$ and $\sigma_i/\sigma_x = 0.82$).

2.3 Atlas

In order to build the atlas, the set of images, I, was first rigidly aligned, using only the first frame, and then non-rigidly registered.

For rigid alignment, the atlas was initialized to a randomly selected instance among the training set, which we denote as A_0, and cropped to completely contain the heart. The rest of the instances were first centred, assuming the mass centre of the epicardium reference segmentation as the centre of the LV, and then rigidly registered to A_0, constraining the transformation to rotations only. The obtained transformations for each of the first frames were extended to the other frames to obtain the registered set IT_0. The intensity average of the images in the IT_0 set was calculated to obtain the first iteration of the atlas, A_1.

For non-rigid alignment, the segmentations of IT_0 were non-rigidly registered to the segmentation of A_1, obtaining the transformation set T_1. The transformations T_1 were then applied to IT_0 and the average of intensities calculated to obtain the atlas, A. Since the segmentation masks were used, convergence was achieved in one single step.

2.4 Initial Segmentation and Rigid Registration

To initialize the rigid alignment, we trained a variation of U-Net [9] for epicardial segmentation. We first down-sampled all the images to 256×256 and normalized them by clipping the smallest and largest 5% intensity values. More details of the network are described in Sect. 2.6. Based on the initial epicardium segmentation of the first frames, we centred and oriented the set of images, I, to the atlas, A, as described in Sect. 2.2.

2.5 Statistical Models of Deformation

We implemented the statistical models of deformation following the SMOD+ method in [8]. Once the rigid registration was completed, the set of segmentations of the images, S, is non-rigidly registered to the atlas segmentation, A_s, obtaining the set of velocity fields, $\{v_i\}$, to diffeomorphically bring each image to the atlas space.

This set $\{v_i\}$ intrinsically encodes the shape variability of the set of images, I, with respect to the reference A. Thus, the distribution of $\{v_i\}$ can be sampled to obtain new velocity fields that implicitly lead to anatomically meaningful deformations within the space of plausible shapes, and we built a statistical model of deformations that can be exploited to generate new images.

In order to generate random deformations, v_g, we first reduced the dimensionality of the distribution of velocity fields by applying principal component analysis (PCA) on the residuals. Then, we sampled the relative weights of the main modes of variation with a multivariate Gaussian distribution, centred at 0 and with standard deviation σ. Finally, each of the images, i, was brought to the atlas space applying v_i and

transformed back to the image space applying the inverse of the random velocity field, v_g. Thus, a new image with the appearance of image i but a random shape within the space of variability of the original images was obtained.

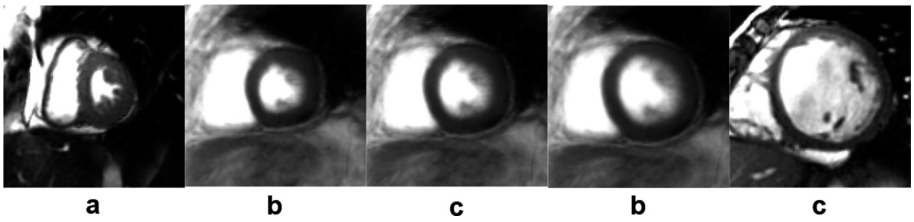

Fig. 1. Atlas and extreme cases of LV shape. The images were input images of the neural networks with the size of 128 × 128 and pixel-spacing of 1.1 mm. (*a*) and (*e*) are the smallest and largest LVs from the original dataset, respectively; (*b*) to (*d*) are generated by PCA mode 1 with $\sigma = -3$, $\sigma = 0$ and $\sigma = +3$.

2.6 Augmentation Strategies

We implemented two augmentations strategies: (1) standard augmentation based on random flipping, rotations (0–360°) and translations (±11 mm in x and y); and (2) augmentation based on SMOD+, which we combined with standard augmentation samples due to the large variability of LV sizes shown in Fig. 1.

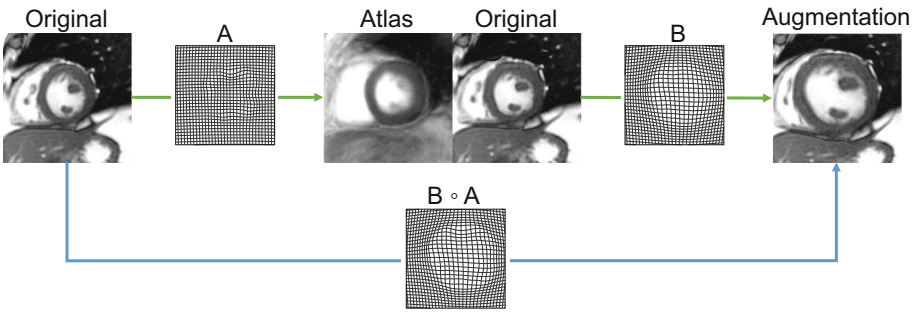

Fig. 2. Combined transformation stages.

The different transformations needed to generate a new image were mathematically combined by convolution as shown in Fig. 2, and therefore the images were interpolated and resized only once at the end of the process. The final resolution used as the neural network input was 128 × 128, with a pixel-spacing of 1.1 mm.

2.7 Neural Network

We compared two neural networks for the final segmentation task, which were a variation of U-Net [9] and a segmentation network based on VGG-16 [6]. For both networks, there were four 2 × 2 max-pooling stages with the stride of 2 × 2, and the number of filters were 64, 128, 256, 512 and 1024 for each stage accordingly. The size of all kernels was 3 × 3 and the activation functions were ReLU for all layers other than the output layer, which was sigmoid. The key difference between the networks was the up-sampling process, with step-by-step up-sampling stages for the U-Net and concatenated up-samples from each scale for VGG-16. This difference is shown in Fig. 3. We implemented the training with cross-entropy as the loss function and Adadelta as optimizer.

The initial segmentation network introduced in Sect. 2.3 shared the same U-Net architecture, while the input size was 256 × 256 and the number of filters were decreased to 16, 32, 64, 128 and 256 for efficiency.

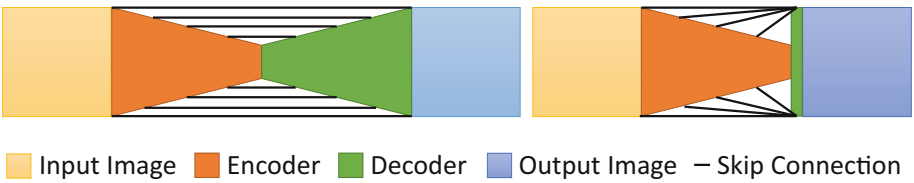

Fig. 3. Schematics of the two neural networks compared in this paper.

2.8 Metrics Evaluation

In the absence of a detailed description of metrics calculation in the challenge, the following approach was adopted. Metrics were calculated from our segmentation results by first converting the neural network outputs to binary masks and then thresholding at 0.5. We extracted the largest object from the binary masks and filled any existing holes. The areas were calculated by multiplying the pixel area times the number of pixels of the region. The dimensions were calculated by averaging the distances between the endocardial contour points and the cavity centroid within the corresponding section. To calculate the regional wall thickness, we first calculated the middle contour of the myocardium and then averaged the closest point-to-point distances between both endocardial and epicardial contours to the middle contour for each middle contour point. The phase estimation was calculated by first defining the frames with maximum and minimum cavity areas to be end-diastolic (ED) and end-systolic (ES) frames, and then assigning linearly interpolated labels to the other frames.

Applying our metrics estimation method to the reference segmentations provided by challenge organisers led to a bias when compared to the set of reference metric values also provided by challenge organisers. Such differences with respect to the provided dataset introduced unnecessary complexity when designing the pipeline and could have been at least partly (and for the areas fully) eliminated with the provision of detailed descriptions of metrics calculation by the organisers. To minimize possible

errors that could be introduced within this stage, we calculated a correction factor λ using the reference area metrics (Ar) and the segmentation estimated areas (As), which was the only metric independent of LV orientation, by minimizing the error of ($Ar - \lambda As$). The square-root of λ was then multiplied to 1D metrics estimated from the segmentation.

3 Experiments and Results

We performed 3-fold cross-validation experiments on the training dataset, with the size of each fold being 18, 19 and 19. The subjects were randomly assigned to one-fold, and for each cross-validation experiment we used 4 subjects as validation set and kept the rest as training set.

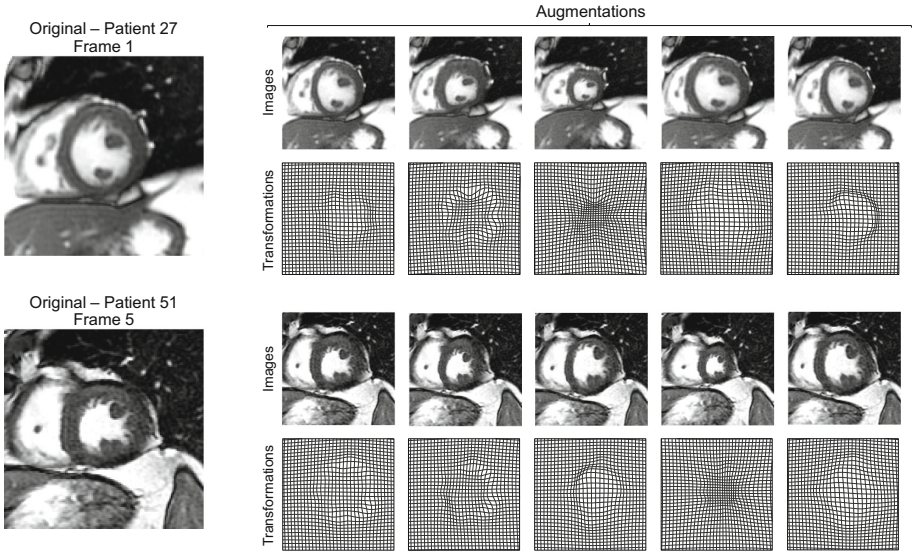

Fig. 4. Example of generated augmentations. Five randomly generated augmentations are shown for each of the two images. The augmented cases varied in size, shape and myocardium thickness of LV.

A model of deformation (described in Sect. 2.4) was learnt for each fold, and the metrics correction factors (described in Sect. 2.7) were also calculated for each fold independently. The network parameters were updated using the training set and model selection was performed using the validation set with early stop. For each training epoch, new instances of training images were randomly generated and used to update the network parameters. Examples of resultant augmented images were shown in Fig. 4, along with the combined transformations.

Table 1. Mean absolute error results.

	Base-line	U-Net Std.	U-Net σ = 1	U-Net σ = 2	U-Net σ = 3	VGG Std.	VGG σ = 1	VGG σ = 2	VGG σ = 3	Test
Dice Endo	1 ±0	0.950 ±0.02	0.949 ±0.033	**0.953** ±0.023	0.950 ±0.026	0.938 ±0.045	0.941 ±0.031	0.943 ±0.032	0.941 ±0.031	N/A
Dice Epi	1 ±0	0.966 ±0.01	0.965 ±0.019	**0.967** ±0.014	0.965 ±0.013	0.957 ±0.026	0.958 ±0.020	0.959 ±0.021	0.959 ±0.020	N/A
A1 (mm^2)	24 ±19	102 ±87	107 ±111	92 ±83	100 ±83	101 ±121	95 ±77	90 ±79	101 ±84	184
A2 (mm^2)	25 ±19	132 ±105	142 ±105	121 ±98	135 ±105	140 ±121	155 ±131	126 ±110	140 ±112	525
Areas (mm^2)	25 ±19	117 ±97	125 ±109	**106** ±91	118 ±96	121 ±122	125 ±111	108 ±97	120 ±101	355
Dim1 (*mm*)	0.64 ±0.72	1.59 ±1.53	1.73 ±2.30	1.46 ±1.28	1.58 ±1.45	1.40 ±1.78	1.40 ±1.16	1.32 ±1.08	1.51 ±1.23	2.59
Dim2 (*mm*)	0.68 ±0.74	1.70 ±1.39	1.81 ± 2.18	1.53 ± 1.31	1.60 ± 1.38	2.17 ± 3.01	1.72 ± 1.48	1.89 ± 1.65	1.96 ±1.77	2.33
Dim3 (*mm*)	0.77 ±0.87	1.65 ± 1.32	1.63 ± 1.59	1.56 ± 1.28	1.64 ± 1.26	2.01 ± 2.54	1.76 ± 1.55	1.83 ± 1.62	1.81 ±1.48	2.40
Dims (*mm*)	0.69 ±0.78	1.65 ±1.42	1.72 ±2.05	**1.52** ±1.29	1.60 ±1.36	1.86 ±2.52	1.63 ±1.42	1.68 ±1.49	1.76 ±1.52	2.44
RWT1 (*mm*)	0.35 ±0.45	1.01 ±1.03	0.98 ±0.89	0.85 ±0.68	0.89 ±0.68	0.89 ±0.75	1.01 ±0.89	0.91 ±0.80	0.91 ±0.84	2.40
RWT2 (*mm*)	0.41 ±0.37	1.23 ±0.85	1.23 ±0.92	1.05 ±0.78	1.18 ±0.86	1.19 ±0.90	1.19 ±0.94	1.15 ±0.84	1.22 ±0.87	2.39
RWT3 (*mm*)	0.33 ±0.27	1.27 ±0.97	1.22 ±0.95	1.10 ±0.87	1.26 ±1.03	1.21 ±0.97	1.21 ±0.97	1.15 ±0.92	1.26 ±1.00	2.20
RWT4 (*mm*)	0.36 ±0.45	1.20 ±0.90	1.27 ±0.97	1.21 ±1.02	1.23 ±1.00	1.22 ±0.99	1.16 ±0.93	1.13 ±0.91	1.29 ±1.10	1.91
RWT5 (*mm*)	0.41 ±0.37	0.91 ±0.75	1.02 ±0.79	1.00 ±0.74	0.97 ±0.78	1.10 ±1.30	0.93 ±0.78	0.99 ±0.89	1.14 ±0.97	1.98
RWT6 (*mm*)	0.45 ±0.43	0.91 ±0.76	0.85 ±0.72	0.84 ±0.66	0.88 ±0.67	1.08 ±1.31	1.02 ±0.78	0.99 ±0.92	0.93 ±0.74	2.21
RWT (*mm*)	0.38 ±0.40	1.09 ±0.9	1.10 ±0.89	**1.01** ±0.81	1.07 ±0.86	1.12 ±1.06	1.09 ±0.89	1.06 ±0.89	1.12 ±0.94	2.18
Phase (%)	2.0	7.9	8.3	8.0	8.1	8.2	**7.8**	8.4	8.2	9.5

Results of the experiments are shown in Table 1. Errors in LV metrics obtained from ideal segmentations are reported in the baseline experiment, which used the reference segmentation provided by challenge organizers after applying the correction factor. For area metrics, after applying the correction factor there was still a mean absolute error (MAE) of 25 mm^2, which is around 25% of the MAE with our best segmentation results. Such an error might have been removed shall we had an accurate description of metric calculations. We could also see a 2% phase estimation error, which is purely dependent on cavity areas and introduced during resampling the images, suggesting the reference phase was sensitive to small noise.

Comparing the two networks, the performance of the U-Net was better than VGG-16 based segmentation network for Dice score, area, dimension and regional wall thickness values. Despite a more accurate estimation of the endocardium using the U-Net, VGG-16 achieved a more accurate phase estimation. This could be caused by the effect of noise we detected in the baseline experiments. From the results we could see that there was a negative effect on the segmentation task by removing multiple stage up-sampling, even though VGG-16 is deeper in the down-sampling stages.

Comparing the two augmentation strategies, our modified SMOD+ approach with $\sigma = 2$ produced the best results. The performance of $\sigma = 1$ and $\sigma = 3$ were limited because the variation of the deformation was either too close to the atlas or far enough to become unrealistic, and for both cases the generalization of the network was disrupted with either unbalanced data or unexpected data. By calculating the p-values, we found significant differences between the two augmentation strategies.

Bland-Altman plots were produced to show the agreement between our best performing network with the reference metrics in Fig. 5. The vast majority of the data points lies within $mean \pm 1.96 \times std$ suggesting a good agreement between the two measurements.

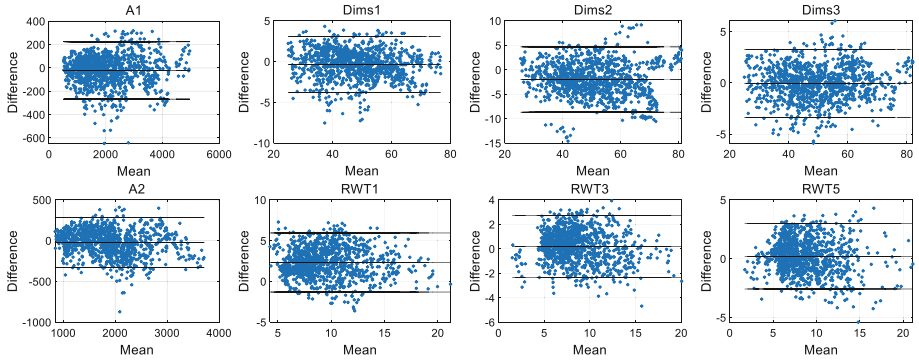

Fig. 5. Bland-Altman plots for U-Net with σ = 2. Areas, dimensions and three regional wall thickness metrics are shown.

We also evaluated qualitatively the segmentation results of our best performing network. Three examples with Dice score from high to low (including the worst case) are shown in Fig. 6. Our segmentation results from neural networks appeared to be consistent with image features, however, the manual reference segmentation contours were comparatively independent from image features. The values of Dice score showed similarity between our segmentation results and the provided references, and larger Dice score represented better similarity between the two.

For the testing dataset we used the entire training dataset to get the model of deformation and the correction factor for better generalization. We used all three networks of U-Net with $\sigma = 2$ and embedded the neural network predictions by averaging before calculating the metrics. The performance of our pipeline on the testing dataset is shown in Table 1. Comparing to the cross-validation experiment results using

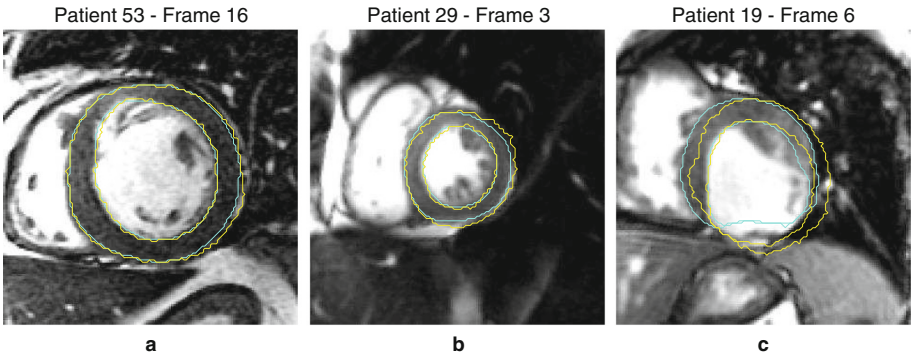

Fig. 6. Segmentation results of training dataset. (*a*) to (*c*) correspond to Dice scores from high to low. The yellow contour is the reference segmentation, and the cyan contour is our proposed U-Net result based on SMOD+ augmentation. (Color figure online)

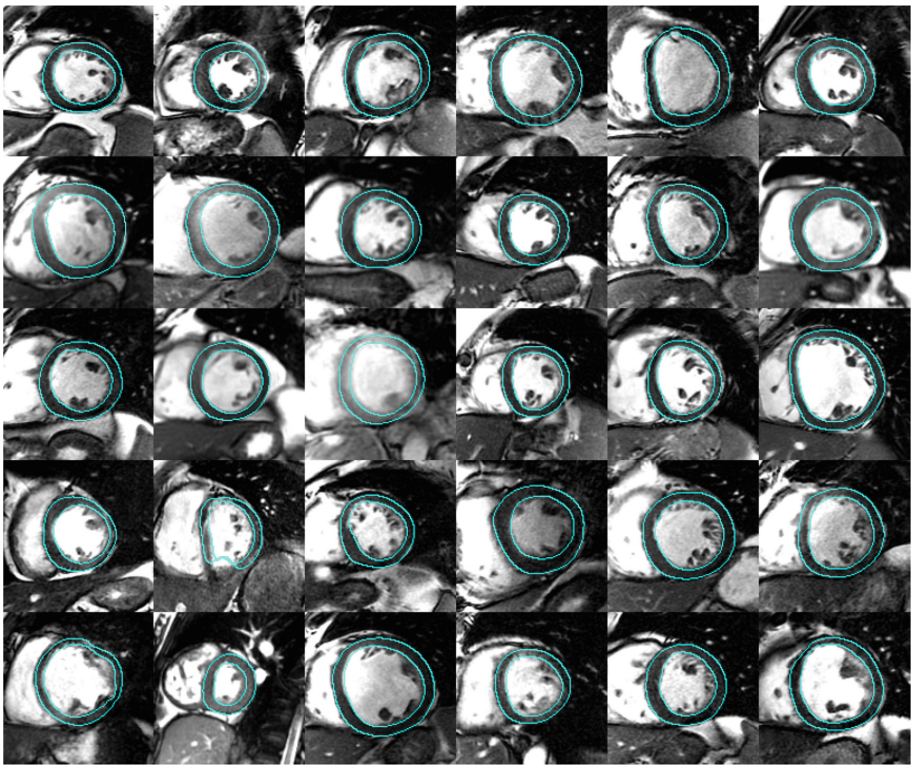

Fig. 7. Segmentation results of the testing dataset. We presented the segmentation result of the first frame of all subjects within the testing dataset. Different from the training dataset, there were no reference segmentation provided, and therefore only the segmentation results from our proposed neural networks are shown.

the training dataset, the testing dataset errors are comparatively larger. For metric A2 (representing myocardium area), the testing result mean absolute error is more than 4 times bigger and reaches 525 mm^2, which is larger than a square with the side of 2 cm.

In order to further investigate this difference between training dataset and testing dataset results, we produced the segmentation result of all the subjects in the testing data for qualitative analysis. Results for the first frame of each subject are shown in Fig. 7. From the visual inspection, the segmentation results of the testing dataset were comparable with the training dataset, and there was no clear evidence suggesting why would the metric evaluation of the testing dataset performed worse than in the training dataset based on the segmentation results.

4 Conclusion

In this paper, we have proposed a full quantification pipeline of the LV using CMR mages, developed and applied to the MICCAI 2019 Left Ventricle Full Quantification Challenge. We performed 3-fold cross-validation experiments on the training dataset, and for all the combinations of network structure and augmentation strategies, U-Net with our modified SMOD+ augmentation achieved the best results within our pipeline, showing the benefits of using multi-stage up-sampling and advanced augmentation strategies.

Compared to MICCAI 2018 Left Ventricle Full Quantification Challenge, the dataset was closer to real-life clinical conditions by removing the pre-processing of the images. At the same time, the size of the training dataset was reduced from 145 to 56 subjects. Both changes made the task significantly more challenging, which steered our focus towards the pre-processing and metrics evaluation stages, as well as the implementation of an anatomically meaningful augmentation method to enhance the neural network performance. Despite the more challenging task, our method achieved comparable results to last year's participants for both cross-validation on the training dataset and the final testing dataset.

The performance of our pipeline on the testing dataset did not reach the level of our cross-validation experiments, and based on the provided qualitative evaluation of the segmentation results the reason of such big differences between the mean absolute errors remains unclear to us. Similar performance drops in testing datasets were also identified in all the best ranking methods in MICCAI 2018 Left Ventricle Full Quantification Challenge [1–4]. It appears to us that this phenomenon is less dependent on the candidate methods, but rather closely related to the distribution of subjects in the training and testing dataset. Additional details on the testing dataset, and an explicit description of metrics calculation, would facilitate the interpretability of these results and improve future challenges.

Acknowledgments. This work was supported by the European Unions Horizon 2020 research and innovation program under the Marie Sklodowska-Curie (g.a. 764738) and by the British Heart Foundation (PG/16/75/32383). Authors are financially supported by a Wellcome Trust Senior Research Fellowship (to PL, 209450/Z/17/Z) and a BHF Intermediate Basic Science Research Fellowship (to ABO, FS/17/22/32644).

References

1. Li, J., Hu, Z.: Left ventricle full quantification using deep layer aggregation based multitask relationship learning. In: Pop, M., et al. (eds.) STACOM 2018. LNCS, vol. 11395, pp. 381–388. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-12029-0_41
2. Kerfoot, E., Clough, J., Oksuz, I., Lee, J., King, A.P., Schnabel, J.A.: Left-ventricle quantification using residual U-Net. In: Pop, M., et al. (eds.) STACOM 2018. LNCS, vol. 11395, pp. 371–380. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-12029-0_40
3. Xu, H., Schneider, J.E., Grau, V.: Calculation of anatomical and functional metrics using deep learning in cardiac MRI: comparison between direct and segmentation-based estimation. In: Pop, M., et al. (eds.) STACOM 2018. LNCS, vol. 11395, pp. 402–411. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-12029-0_43
4. Guo, F., Ng, M., Wright, G.: Cardiac MRI left ventricle segmentation and quantification: a framework combining U-Net and continuous max-flow. In: Pop, M. (ed.) STACOM 2018. LNCS, vol. 11395, pp. 450–458. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-12029-0_48
5. UK Biobank. https://www.ukbiobank.ac.uk/
6. Bai, W., et al.: Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. J. Cardiovasc. Magn. Reson. **20**(1), 65 (2018)
7. Petersen, S.E., et al.: The impact of cardiovascular risk factors on cardiac structure and function: Insights from the UK Biobank imaging enhancement study. PLoS ONE **12**(10), e0185114 (2017)
8. Corral Acero, J., et al.: SMOD - data augmentation based on statistical models of deformation to enhance segmentation in 2D cine cardiac MRI. In: Coudière, Y., Ozenne, V., Vigmond, E., Zemzemi, N. (eds.) FIMH 2019. LNCS, vol. 11504, pp. 361–369. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-21949-9_39
9. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
10. Xue, W., Brahm, G., Pandey, S., Leung, S., Li, S.: Full left ventricle quantification via deep multitask relationships learning. Med. Image Anal. **43**, 54–65 (2018)
11. Orbes-Arteaga, M., et al.: PADDIT: probabilistic augmentation of data using diffeomorphic image transformation. In: Medical Imaging 2019: Image Processing, vol. 10949, p. 109490S. International Society for Optics and Photonics, March 2019