Left Ventricle Quantification Challenge: A Comprehensive Comparison and Evaluation of Segmentation and Regression for Mid-ventricular Short-axis Cardiac MR Data

Wufeng Xue, Jiahui Li, Zhiqiang Hu, Eric Kerfoot, James Clough, Ilkay Oksuz, Hao Xu, Vicente Grau, Fumin Guo, Matthew Ng, et al.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3064353, IEEE Journal of Biomedical and Health Informatics

Left Ventricle Quantification Challenge: A Comprehensive Comparison and Evaluation of Segmentation and Regression for Mid-ventricular Short-axis Cardiac MR Data

Wufeng Xue, Jiahui Li, Zhiqiang Hu, Eric Kerfoot, James Clough, Ilkay Oksuz, Hao Xu, Vicente Grau, Fumin Guo, Matthew Ng, Xiang Li, Quanzheng Li, Lihong Liu, Jin Ma, Elias Grinias, Georgios Tziritas, Wenjun Yan, Angélica Atehortúa, Mireille Garreau, Yeonggul Jang, Alejandro Debus, Enzo Ferrante, Guanyu Yang, Tiancong Hua, and Shuo Li*

***Abstract*—Automatic quantification of the left ventricle (LV) from cardiac magnetic resonance (CMR) images plays an important role in making the diagnosis procedure efficient, reliable, and alleviating the laborious reading work for physicians. Considerable efforts have been devoted to LV quantification using different strategies that include segmentation-based (SG) methods and the recent direct regression (DR) methods. Although both SG and DR methods have obtained great success for the task, a systematic platform to benchmark them remains absent because of differences in label information during model learning.**

In this paper, we conducted an unbiased evaluation and comparison of cardiac LV quantification methods that were submitted to the Left Ventricle Quantification (LVQuan) challenge, which was held in conjunction with the Statistical Atlases and Computational Modeling of the Heart (STACOM) workshop at the MICCAI 2018. The challenge was targeted at the quantification of 1) areas of LV cavity and myocardium, 2) dimensions of the LV cavity, 3) regional wall thicknesses (RWT), and 4) the cardiac phase, from mid-ventricle short-axis CMR images. First, we constructed a public quantification dataset Cardiac-DIG with ground truth labels for both the myocardium mask and these quantification targets across the entire cardiac cycle. Then, the key techniques employed by each submission were described. Next, quantitative validation of these submissions were conducted with the constructed dataset. The evaluation results revealed that both SG and DR methods can offer good LV quantification performance, even though DR methods do not require densely labeled masks for supervision. Among the 12 submissions, the DR method LDAMT offered the best performance, with a mean estimation error of 301 mm^2 for the two areas, 2.15 mm for the cavity dimensions, 2.03 mm for RWTs, and a 9.5% error rate for the cardiac phase classification. Three of the SG methods also delivered comparable performances. Finally, we discussed the advantages and disadvantages of SG and DR methods, as well as the unsolved problems in automatic cardiac quantification for clinical practice applications.

***Index Terms*—left ventricle, quantification, segmentation, regression, deep neural network**

Wufeng Xue is with the National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China, and also with the Department of Medical Imaging, Western University, London, ON N6A 3K7, Canada. Shuo Li is with with the Department of Medical Imaging, Western University, London, ON N6A 3K7, Canada. Jiahui Li is with Beijing University of Post and Telecommunication, Beijing, China Zhiqiang Hu is with Peking University, Beijing, China. Eric Kerfoot, James Clough and Ilkay OKsuz are with School of Biomedical Engineering & Imaging Sciences, King's College London, UK. Hao Xu and Vicente Grau are with Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK. Fumin Guo and Matthew Ng are with Sunnybrook Research Institute, Department of Medical Biophysics, University of Toronto, Canada. Xiang Li and Quanzheng Li are with Department of Radiology, Massachusetts General Hospital, Boston, USA. Lihong Liu and Jin Ma are with with Pingan Technology (Shenzhen) Co.Ltd. Elias Grinias and Georgios Tziritas are with Department of Computer Science, University of Crete, Heraklion, Greece. Wenjun Yan is with Department of Electrical Engineering, Fudan University, Shanghai, China. Angélica Atehortúa is with LTSI UMR 1099, F-35000 Rennes, France and also with Universidad Nacional de Colombia, Bogotá, Colombia. Mireille Garreau is with LTSI UMR 1099, F-35000 Rennes, France. Yeonggul Jang is with Brain Korea 21 PLUS Project for Medical Science, Yonsei University. Alejandro Debus and Enzo Ferrante are with Research Institute for Signals, Systems and Computational Intelligence, sinc(i) , FICH-UNL/CONICET, Santa Fe, Argentina. Guanyu Yang and Tiancong Hua are with Centre de Recherche en Information Biomédicale Sino-Français (CRIBs), Southeast University, Nanjing, China. Guangyu Yang is also with LIST, Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing, China.

* Corresponding author. (E-mail: slishuo@gmail.com)

I. Introduction

CARDIAC disease is one of the leading causes of worldwide morbidity and mortality [1]. As the gold standard of cardiac disease diagnosis, cardiac magnetic resonance (CMR) images have been widely used in routine practice for early detection, decision making, patient management, and treatment evaluation. However, tedious visual inspections and manual delineation have to be conducted by physicians before useful and reliable clinical information can be inferred from the hundreds of images typically used to examine each patient. In addition, the manual contouring of myocardium is typically limited to the end diastolic (ED) and end systolic (ES) frames that makes it insufficient for dynamic function analysis during the entire cardiac cycle. Even in this manner, the obtained results heavily depend on the experience of physicians and often exhibit high inter-observer variation [2].

Considerable efforts have been devoted to CMR image analysis. Specifically, automatizing this procedure by leveraging the techniques from image processing, machine learning, and more recently, deep learning, has been the focus. Accurate

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3064353, IEEE Journal of Biomedical and Health Informatics

quantification of the left ventricle (LV) from cardiac images is one of the most clinically important and frequently demanded tasks for identification and diagnosing of cardiac diseases [3]. To provide a comprehensive global and regional cardiac function assessment, multi-type quantification of cardiac LV is required, which simultaneously quantifies four types of cardiac indices, such as cavity and myocardium areas, regional wall thicknesses, LV dimensions and the cardiac phase, as shown in Fig. 1, for every frame in the entire cardiac cycle. All of these indices are necessary for evaluation of global and regional cardiac function. Detailed definitions and clinical roles of them can be found in [4]–[8] and subsection II-C . Due to the heavy labeling workload of the full stack of short-axis slices, in this challenge we focus on the quantification of one representative mid-ventricle slice as proof of concept.

With regard to cardiac quantification, the following issues must be effectively addressed to achieve reliable and accurate quantification: 1) the appearance of myocardium is difficult to be captured in presence of low contrast structure, inhomogeneity brightness and texture, various pathologies, and high variability of cardiac structure across subjects; 2) the heart motion is a complex non-rigid deformation process that includes regional wall thickening, and circumferential and longitudinal ventricular shortening, thus, it is even more difficult to model.

Two categories of methods exist in LV quantification: segmentation-based (SG) and direct-regression (DR) methods. SG methods, first, intuitively segment the myocardium from its surrounding background structures, and then, they calculate the required cardiac indices from the segmented mask. DR methods circumvent the segmentation procedure and estimate the above-mentioned cardiac indices directly from the image intensities. They build the mapping from image appearance directly to the cardiac indices of interest with the objective of minimizing the quantification error, instead of the segmentation error. Both SG and DR methods have obtained accurate quantification performance with the help of advanced machine learning techniques and manually annotated datasets, which will be respectively detailed below.

A. Existing work on automatic cardiac LV quantification

1) SG methods: Most of the early work [6], [9] on automatic CMR image quantification fall into this category, and were based on the classical image segmentation methods such as level-set, graph-cut, thresholding, and region growing. In some cases, user interaction [10]–[12] and prior information [10], [13]–[15] were required. However, inaccurate prior information and strong user interaction may prevent these methods from efficient clinical application.

Recent works take advantages of the powerful representation capabilities of deep neural networks (DNN) and the rich labeling information of large datasets with densely labeled segmentation masks to train LV segmentation models. Convolution neural networks (CNNs) have achieved great success in cardiac segmentation with optional refinement by deformable model [16] and level set [17], [18]. Fully convolution network (FCN) have been used for cardiac segmentation [19], [20]

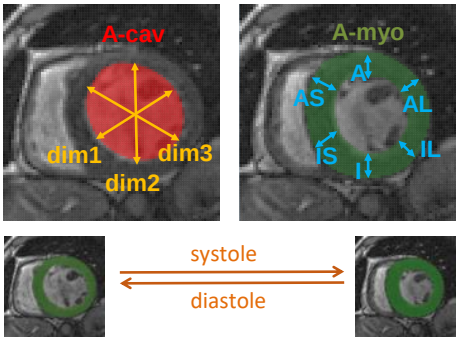

Fig. 1: The four tasks of LV quantification in the LVQuan 2018 challenge: areas of cavity (A-cav) and myocardium (A-myo), directional dimensions of cavity (dim1∼dim3), regional wall thicknesses (for regions of IS, I, IL, AL, A, AS), and cardiac phase (1: systole or 0: diastole). (A: anterior; AS: anterospetal; IS: inferoseptal; I: inferior; IL: inferolateral; AL: anterolateral.)

because of its success in semantic segmentation of natural images [21]. Other CNN structures utilized in cardiac segmentation include 2D/3D Unet [22], grid-like CNN [23], [24], dilated CNN [25], encoder-decoder architecture [26], parallel coarse and fine network in polar space [27], 3D-CNN model with deep supervision [28], and distance map regularized Unet [29]. A comprehensive review of cardiac segmentation can be found in [30]. SG methods can provide not only the quantification results, but also the contour of the myocardium, which can help validate and understand the quantification results.

2) DR methods: When the densely labeled dataset is not available, direct methods without segmentation have grown in popularity in cardiac volume estimation, and obtained effective performance benefiting from machine learning algorithms. The pioneer work along this line followed a two-stage procedure: feature engineering and regression learning. Cardiac images were represented by hand-crafted features [31]–[34] or those obtained by unsupervised learning [35], [36]. Then, cardiac volumes were estimated from these features with a separated regression model. Although these methods demonstrated their effectiveness, they suffer from the vulnerable representation of hand-crafted features and lack of deformation modeling of the heart motion.

Deep learning-based methods leverage the powerful representation learning ability and end-to-end learning framework, thus these limitations are well addressed. CNNs were employed for ROI cropping, slice localization, and volume estimation from cardiac images of ES and ED frames [37]. [38] proposed the first end-to-end framework for cardiac indices quantification based on a cascade convolution auto-encoder and regression network, wherein only the true values of the cardiac indices were utilized to supervise the parameter learning. A two-branch architecture with recurrent neural network (RNN) was proposed in [39] to predict the RWT values of the whole cardiac sequences. More cardiac indices were included in [40] using a multitask neural network setting, where the mutual relatedness within and between tasks were

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3064353, IEEE Journal of Biomedical and Health Informatics

TABLE I: General information of existing datasets, including number of patients, whether the data comes from multiple or single source, age, gender, and representative pathologies. For patient's age, a triple of mean {min, max} is provided for each dataset. For patient's gender, male:female is provided.

	No.	Souce	Age	Gender	Pathology
SCD	45	Multiple	61.1{23,88}	32:13	Heart failure with/without infarction, HCM, and normal
LSVC	200	Multiple	62.73{34,84}	73:22 (training set)	Myocardial infraction
ACDC	150	Single	Unknown	Unknown	Normal, heart failure with infarction, DCM, HCM, abnormal right ventricle
UK BioBank	>20000	Single	56.5{40,69}	54.4% female	Community volunteers
DSB2015	1140	Multiple	42.2{2,88}	669:471	Normal to abnormal cardiac function

modeled by the group lasso and consistency constraints. This task relatedness was later improved by the Bayesian-based relationship learning (DMTRL) [41], which achieved the state-of-the-art performance.

Both SG and DR methods have achieved great success in automated cardiac quantification and provide a great potential for routine clinical application. Therefore, a uniform platform for benchmarking the SG and DR methods simultaneously will help in the advancement of research on cardiac quantification and in accelerating its practical application.

B. Existing datasets for MRI cardiac image analysis

Irrespective of the use of segmentation or direct regression, the ultimate goal of cardiac quantification is to automatically compute the clinically significant cardiac indices which will appear on the report of physicians for reference of further diagnosis, evaluation, and monitoring. Research on cardiac image analysis has been greatly promoted by publicly accessible datasets, particularly those proposed in conjunction with international challenges. However, most of these datasets focus on segmentation of myocardium, whereas a publicly accessible dataset for direct prediction of multiple cardiac indices has rarely been considered. In the following section, we briefly review the existing CMR datasets. Please note that this is not a full coverage of existing CMR datasets.

The Sunnybrook Cardiac Data (SCD) [1] was provided to the public for the cardiac LV segmentation challenge of MICCAI 2009. It consisted of cine-MRI images of 45 subjects from groups of healthy, hypertrophy, heart failure with infarction and heart failure without infraction. Expert-drawn contours of the endocardium and epicardium were provided as the ground truth for segmentation.

Left Ventricular Segmentation Challenge (LVSC) [2] [2] in the STACOM workshop at MICCAI 2011 provided steady-state free precession CMR images in short axis and long axis views from 200 patients with coronary artery disease and regional wall motion abnormalities. Binary masks of the myocardium for the 100 training cases were provided.

The Automatic Cardiac Diagnosis Challenge (ACDC) [3] [42] in MICCAI 2017 provided cine-MRI images of short-axis view, from 150 clinical routine patients, and covers five well-defined groups according to the medical reports: normal, heart failure with infraction, dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), and abnormal RV. Manually-drawn 3D volumes of LV and RV cavities and myocardium at ED and ES frames were provided for segmentation reference.

UK Biobank [43] is currently the largest CMR dataset with more than 20,000 community volunteers and targeting 100,000 subjects. The CMR protocol includes white blood CMR, short- and long-axis cine CMR, strain CMR, flow CMR and parametric CMR. Top quality expert annotations were provided along with the data, including manual delineations of LV endocardium, epicardium, and RV endocardium.

While the previous datasets provide the manual contours or segmentation mask, which require heavy annotation workload. In clinical practice, critical cardiac indices, such as EF and chamber size, are often provided in the physicians' report, and can be directly employed to learn automatic quantification methods, and in the meantime avoid heavy annotation work.

The 2015 Second Annual Data Science Bowl[4] (DSB2015) can be considered as a dataset for direct regression of ejection fraction from CMR images. It provided 2D cine images with approximately 30 frames across the cardiac cycle for a large number of cases (500 for training, 200 for validation and 440 for test), along with the LV volumes at the ED and ES frames. No manually segmented ground truth was provided in this dataset. However, no quantitative information for the frames other than ED and ES frames were provided, thus studies on the dynamic cardiac function analysis cannot be conducted. Besides, apart from EF, other indices as shown in Fig. 1 are also of great clinical significance [4], [5], [7], [8], yet were ignored in this database.

General information of these datasets are also illustrated in Table. I.

C. Contributions

While cardiac quantification has been a hot research topic in medical image analysis and is of great significance in routine clinical practice, there is never a uniform platform that can be employed for benchmarking both SG and DR quantification methods. In this work, we provided such a platform through the Left Ventricle Quantification challenge (LVQuan 2018[5]), which was held in conjunction with the Statistical Atlases and Computational Modeling of the Heart (STACOM) workshop at MICCAI 2018. The main contributions of this work are as follows:

- The LVQuan 2018 Challenge established the first milestone of LV quantification for 2D mid-ventricle short-axis CMR images. The dataset associated with the challenge

[1] http://www.cardiacatlas.org/studies/sunnybrook-cardiac-data/
[2] http://www.cardiacatlas.org/challenges/lv-segmentation-challenge/
[3] https://www.creatis.insa-lyon.fr/Challenge/acdc/
[4] https://www.kaggle.com/c/second-annual-data-science-bowl
[5] https://lvquan18.github.io/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3064353, IEEE Journal of Biomedical and Health Informatics

created a foundation for researches on LV quantification by providing ground truth labels simultaneously for both LV segmentation and multiple cardiac indices.
- The challenge established the state of art for LV quantification methods, revealed the strengths and weaknesses of SG and DR methods and the effects of combining them, discovered the appropriate network architectures for SG and DR methods, and summarized the unsolved problems in cardiac quantification. This will further advance the performance of cardiac quantification and take a step closer to automatic generation of rich diagnostic quantitative reports in practical clinical application.
- The challenge revealed an important but often overlooked fact that without the densely labeled segmentation masks, DR methods can still achieve accurate results that are equally good to those of SG methods. This makes it possible to develop more accurate and stable quantification methods by combining pixel-level supervised SG methods on small dataset with weakly supervised DR methods on large easily acquired dataset.

The remainder of this paper is organized as follows. In Section II, we first present details of the proposed database Cardiac-DIG employed in the challenge. Then, we describe the submissions of the LVQuan 2018 Challenge from all participants, particularly the core techniques utilized in these submissions. The challenge protocols and the evaluation criteria of the quantification performance are described in Section IV. All the methods are analyzed and compared in Section V. Section VI concludes the paper.

II. Cardiac-DIG Dataset

The Cardiac-DIG dataset was developed for the LVQuan 2018 Challenge. The frame-wise labeling of the myocardium mask and multiple indices across the *entire cardiac cycle* enables the dataset to benchmark *both* SG and DR quantification methods with respect to global, regional, and temporal cardiac functions.

A. Data collection

The challenge was held in two stages (i.e., training and test) to ensure a fair procedure. For the training stage, a dataset of 2D cine MR images of 145 subjects was collected from three hospitals affiliated with two health-care centers (London Healthcare Center and St. Joseph's HealthCare). The ages of the subjects were between 16 and 97 years, with average age of 58.9 years. The pixel spacings of the MR images ranged from 0.68 mm/pixel to 2.08 mm/pixel, with a mode of 1.56 mm/pixel. Because all the subjects were collected from clinical practice without any specific selection, pathologies from moderate to severe cardiac issues were present, which included regional wall motion abnormalities, myocardial hypertrophy, dilated cardiomyopathy, mildly enlarged LV, atrial septal defect, LV systolic dysfunction, LAD territory ischemia, and constrictive pericarditis, etc. For the dataset, the LV of each subject was divided into equal thirds perpendicular to the long axis of the heart following the standard AHA prescription [44], and a representative mid-ventricle slice with visible papillary muscle and trabecula was selected. Although quantification of full stack of short-axis slices leads to accurate calculation of LV volumes, it also results in remarkable heavy labeling work. This challenge focused on the mid-cavity slices. Each subject contained 20 frames of the mid-ventricle slice throughout a whole cardiac cycle, which was obtained with electrocardiogram-gating and breadth-holding.

For the test stage, another dataset that included the 2D short axis cine MR images of 30 subjects from the same institutions, was collected in the same manner as in the training dataset. Demographic information of the training and test datasets can be found in Table II.

B. Pre-processing

All cardiac images in both the training and test datasets underwent several preprocessing steps, as shown in Fig. 2, including 1) landmark labeling, wherein the two intersections of LV and RV are manually marked as the reference points; 2) rotation, which makes the line that connects the two landmarks vertical; 3) ROI cropping, wherein a square area encloses the LV with size twice the distance between the two landmarks; and 4) resizing, wherein all the ROI images are resized to a standard size of 80×80. In this procedure, the same rotation, cropping, and resizing were applied to all the frames of a patient to preserve the original myocardium motion across the cardiac cycle. After this procedure, the images from different patients are approximately aligned in size, orientation, and scale.

The pre-processing procedure can 1) alleviate the difficulties of the task a lot and make reserchers focus on the quantification or segmentation of the myocardium, and 2) make the evaluation not biased by various pre-processing.

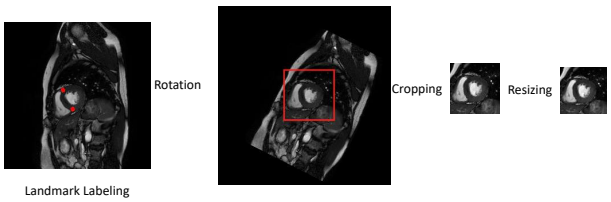

Fig. 2: Procedure of the pre-processing, which includes landmarking labeling, rotation, cropping, and resizing to normalize images of all subjects in size, orientation, and scale.

C. Ground truth

We further provided the ground truth values of the aforementioned clinical related measurements (as shown in Fig. 1) for the whole MR sequences of our dataset. These information are closely related to heart failure, cardiomyopathy, myocardial infarction and hypertension [6] and are critical for computation of recommended quantitative information (ejection fraction (EF), LV mass, LV volumes) according to the clinical guidelines [4], [5], [45]. These measurements are illustrated in Fig. 1 and described as follows:

- Areas of cavity and myocardium (mm^2), which describe the size of the blood pool and myocardium. With areas of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3064353, IEEE Journal of Biomedical and Health Informatics

the full stack SAX slices, we could obtain LV volumes, LV mass and EF, which are essential for cardiac function evaluation and cardiac events monitoring [46].

- Directional dimensions (mm), which describe the size of the LV cavity and can be used to estimate the LV volume, and together with the wall thicknesses, to categorize the hypertrophy and remodeling [5]. The three dimensions are in directions of AS-IL, IS-AL and A-I.
- Regional wall thicknesses (mm), which describe the mean thickness of the myocardium in each segment and can be employed to quantify regional dysfunction, such as those seen in myocardial ischemia or after myocardial infarction [6].
- Cardiac phase (0/1), which is a binary vector and indicates whether a frame is diastolic or systolic (0/1) across the whole sequences.

For each subject, all frames in the cardiac sequence were manually contoured to obtain the epicardial and endocardial borders, which were double-checked by two experienced physicians each having more than 10 years of experience in diagnostic imaging. The LV cavity and myocardium areas can be easily obtained by counting the pixel numbers in the binary masks of the cavity and myocardium. The regional wall thicknesses were obtained by using a center line method. First, myocardial thicknesses were automatically acquired from the two borders in 60 measurements using the 2D centerline method [47]. Then, the myocardium was divided into six segments (as shown in Fig. 4 of [44]), with 10 measurements per segment. Finally, these measurements were averaged per segment as the ground truth of regional wall thicknesses. The cavity dimension for each direction was calculated by averaging the distance of 10 pairs of points that from the opposite positions of the endocardial border. Papillary muscles and trabeculations were excluded in the myocardium. The values of RWT and cavity dimensions were normalized by the image dimension, whereas the areas were normalized by the total number of pixels. During evaluation, the obtained results were converted to physical thickness (in mm) and area (in mm^2) by reversing the resizing procedure and multiplying the pixel spacing for each subject. The cardiac phase was obtained based on whether the cavity area increases or decreases for two successive frames.

It is worth noting that we aim to quantify the LV, not only for the ES and ED frames (as in most existing segmentation methods) but also for frames across the entire cardiac cycle. In this way, more complex cardiac functions can be explored such as wall thickening, cardiac remodeling. The statistical information of these measurements for the training and test datasets can be found in Table II. We can draw that all the measurements are roughly aligned, with the supports for the training dataset being slightly larger.

III. Submissions to the LVQuan18 Challenge

The challenge was launched in November 27, 2017, when the training data were released and accessible on request. The challenge attracted wide interests from institutions around the world, and a total of 49 requests of dataset were received. Among them, 12 teams successfully submitted a paper describing their algorithm, and test results for final performance evaluation. These submissions fell into three categories: SG, DR, and Combined (COM). These methods employed various techniques trying to solve the challenges of the quantification task mentioned previously. Specifically, CNN and encoder-decoder networks were mostly utilized to capture image features effectively at various levels, whereas RNN and 3D CNN were employed to model the temporal dynamic dependency during the deformation procedure of the cardiac cycle. Besides, the multi-task relationship [41] was also modeled in the DR methods to deliver compatible multiple predictions. Fig. 3 demonstrates the conceptual architectures of some key networks. In the following, we describe the details of all submitted methods. The original papers can be accessed in [48].

TABLE II: Demographic information and the statistics of the ground truth labels in the training/test dataset. For each continuous variable, a triple of {min, median, max} is provided. For cardiac phase, numbers of the two classes are displayed.

	Training dataset	Test dataset
Sex	Male: 50 Female: 20 Unknown:73	Male: 20 Female:10
Age (year)	{16, 59, 97}	{11, 58, 80}
Weight (kg)	{45.36, 86.18, 230}	{70, 79.5, 93}
EF (%)	{9.6, 40.6, 73.8}	{16, 42.5, 87}
A-cav (mm^2)	{485, 2099, 4936}	{535, 2118, 4560}
A-myo (mm^2)	{788, 1922, 3812}	{1164, 2453, 4181}
Dimensions (mm)	{23.8, 51.5, 81.0}	{24.2, 52.0, 78.9}
RWTs (mm)	{1.40, 8.76, 24.43}	{4.67, 10.96, 22.86}
Phase	Diastolic: 1680 systolic: 1220	Diastolic: 342 systolic: 258

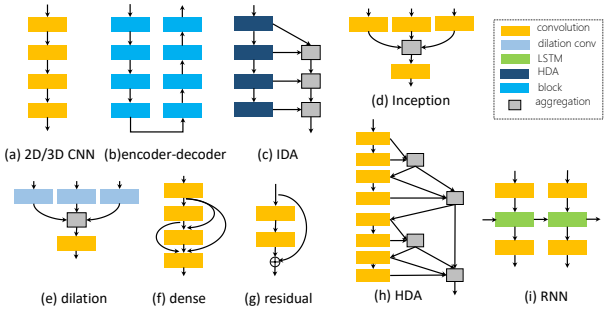

Fig. 3: Various network structures are utilized by the submissions, to extract powerful cardiac representations (a, b), to enhance the feature extractions (d, e, f, g), to aggregate features from multi-scale and resolutions (c, h), and to model the temporal dynamics of cardiac motion (i).

A. SG methods

ResUNet [49] tackled the LV quantification task as myocardium segmentation and employed the well-know segmentation network Unet [50] (encoder-decoder like architecture, as shown in Fig. 3(b)) in a modified manner, in which the convolution layers in each block were replaced by residual units (Fig. 3(g)). The residual unit has been shown to improve the diversity of network depth, thus leading to improved performance over the traditional convolution layer. During net-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3064353, IEEE Journal of Biomedical and Health Informatics

work training, free-form deformation-based data augmentation was employed to improve the generalization ability. Finally the cardiac indices were calculated by projecting the segmentation mask into the polar space.

ESUPNet [51] proposed a cascading DNN for LV quantification. Again the Unet (Fig. 3(b)) was employed for segmentation, with inception-like units inserted after the first and second blocks of the contracting path, for the purpose of integrating information of multi-scales and enhancing high-level feature learning and supervision. Then, an RNN (Fig. 3(i)) for temporal dynamic modeling was used for the phase classification from the LV indices that were obtained from the segmentation result.

UNetMF [52] made use of the Unet (Fig. 3(b)) to generate a probability map, which was then used to initialize a continuous Max-flow segmentation for refinement. Image edge information was used to regularize the final segmentation result.

SegNetRF [53] first obtained high-level features using a semantic segmentation method SegNet [54], which has a similar structure to the encoder-decoder (Fig. 3(b)). The network was initialized with weights of an ImageNet-pretrained VGG-16. To retrieve the two contours in a reliable manner, a series of refinement procedures were followed, which included myocardium and cavity refinement, ED localization, ED refinement based on salient perceptual grouping model, and temporal refinement.

CNTCVX [55] contributed the only non-DNN-based method, which was an unsupervised image-driven method for LV segmentation. The entire pipeline included intensity-based image analysis for cavity localization, statistic-based feature extraction for the myocardium, cavity, and background, and then prior knowledge based constraints such as *connectivity* and near *convexity*.

B. DR methods

LDAMT [56] made use of a deep layer aggregation (LDA) network [57] to fuse information from different layers and scales to directly predict the cardiac indices for the three regression task. Specifically, the LDA network introduced two novel structures, namely, iterative deep aggregation (IDA, Fig. 3(c)) and hierarchical deep aggregation (HDA, Fig. 3(h)), to better fuse semantic and spatial information. The multi-task relationship regularization proposed in [41] was used during network training to ensure the consistency of multiple outputs. The cardiac phase was then predicted from a polynomial fitted to the cavity area to avoid prediction noise.

HQNet [58] followed the framework of [41] and proposed a multi-task quantification network (HQNet). The network was constituted by a hierarchical 3D multi-scale convolution neural network (HCNN) for feature extraction and two LSTM networks for temporal modeling (Fig. 3(i)). Inception-like modules (Fig. 3(d)) with 3D convolution was used in the HCNN. In addition, multi-task relationship constraints [40], [41] were also utilized in the objective function to improve the final estimation accuracy.

CNN3DST [59] proposed a CNN that consisted of an encoder-CNN (Fig. 3(a)) for feature extraction and a spatial-temporal CNN for temporal dynamics modeling and fusion of spatial and temporal information. Two branches for the regression tasks and the classification yielded the final predictions. In the network, a stack of k adjacent frames were taken as input to predict the corresponding indices of the center frame.

FNN2D3D [60] used an FCN (Fig. 3(a)) for feature extraction, where the authors introduced the module of alternative 3D and 2D convolutions to utilize the temporal information. Two parallel paths were then followed for the regression and the classification tasks, respectively.

Besides these DR methods in the challenge, we also include one of the pioneer work DMTRL [41] for performance comparison. DMTRL leverages the powerful representation of deep neural network and learns mutual relationship between different tasks to improve the generalization ability of the learning model.

C. Combined methods

MMED [61] proposed a collection of LV quantification methods under a common network architecture with multiple modes, in order to conduct a fair comparison of the SG and DR methods and to provide a basic idea of the degree to which the binary masks can assist the quantification task. The network architecture included the frequently used encoder-decoder (Fig. 3(b)) for feature extraction, and then two flexible branches for segmentation and regression. The whole network could work in the mode of SG, DR, or combined. The role of temporal dynamic modeling was also studied using an RNN module between the encoder and decoder. The three different modes of the method were denoted as **MMED-S**, **MMED-R**, and **MMED-SR**, respectively.

EnCNNU [62] presented an ensemble learning method for the LV quantification task. Specifically, it leveraged complementary information from the CNN-based (Fig. 3(a)) direct regression and the Unet-based (Fig. 3(b)) segmentation, by ensemble learning with a gradient boosting algorithm.

DenseUMT [63] combined information of both segmentation mask and quantification label to help train the network. Unet (Fig. 3(b)) for segmentation was first enhanced using dense blocks (Fig. 3(f)) and dilation blocks (Fig. 3(e)), and then the index-specific feature was extracted using a shallow CNN. This was followed by two RNN modules (Fig. 3(i)) for temporal dynamic modeling.

IV. Performance Evaluation

A. Configurations

The challenge included a training phase and a test phase. During the training phase, both the cardiac images and their labels (i.e, the segmentation mask and the values of all cardiac indices) for 145 subjects were released to all participants. Five-fold cross validation (CV) was recommended during this phase for model validation. No external data were allowed in this phase (Network weights pre-trained with ImageNet were granted because of the appearance difference between natural and medical images). During the test phase, only the cardiac images of the 30 subjects were provided to the participants according to the challenge schedule. The labels of segmentation masks and quantification values were kept by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3064353, IEEE Journal of Biomedical and Health Informatics

TABLE III: Performance of all submissions for the LVQuan 2018 challenge on the training dataset with 145 subjects under the *five-fold* cross validation (CV) protocols. For each task, only the average performance is shown here and the best result is highlighted in boldface. Average MAE is shown for areas, dimensions and RWTs, and error rate is shown for cardiac phase. All the methods achieved performance better or close to the state-of-the-art DMTRL.

Methods	Area (mm^2)	Dimension (mm)	RWT (mm)	Phase (%)
SG-based methods				
ResUNet	62.3	**0.79**	**0.68**	**6.72**
ESUPNet	**62**	1.14	0.96	8
UNetMF⋆	141.7	1.77	1.39	-
MMED-S	120	1.25	1.03	7.8
SegNetRF†	-	-	-	10
CNTCVX	176	2.23	1.75	10.3
DR-based methods				
LDAMT	**156**	**2.03**	**1.38**	8.1
MMED-R	158	2.08	1.51	9.4
HQNet*	197	2.57	1.51	9.8
CNN3DST	190	2.29	1.42	**3.85**
FNN2D3D	188	2.42	1.42	8.76
Combined methods				
MMED-SR	142	**1.94**	1.42	8.6
EnCNNU‡	**124**	2.27	1.62	13.7
DenseUMT	173	2.44	**1.37**	**7.8**
DMTRL	180	2.51	1.39	8.2

⋆: 45 subjects in the training set were used for test; †: only Pearson correlation coefficients were report on the training set; *:7-fold CV was used; ‡:3-fold CV was used. For MMED, the results for all three modes were reported.

the organizer to conduct the final performance evaluation of different submissions.

B. Evaluation Criteria

The performances of all submissions were evaluated in terms of estimation accuracy of all frames in the cardiac cycle. For the three types of LV indices, the mean absolute error (MAE) between the ground truth and the estimated values were computed to assess the estimation accuracy. For the cardiac phase, error rate (ER) was computed. Let $y \in \mathcal{R}^N$ and $\hat{y} \in \mathcal{R}^N$ be the two vectors of the predicted values and the true values of one cardiac index, where N is the total image number. These evaluation criteria were computed as:

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N} |y^i - \hat{y}^i|, \tag{1}$$

$$\mathrm{ER}(y, \hat{y}) = \frac{\sum_{i=1}^{N} \mathbf{1}(y^i \neq \hat{y}^i)}{N} 100\%. \tag{2}$$

V. Results and Analysis

All three categories of methods for cardiac quantification were evaluated within the same framework, 1) to examine their capability of providing accurate quantification results for the task of LV quantification and their generalization ability to novel data, 2) to compare the performance between SG and DR methods and reveal their strengths and weakness, and 3) to reveal problems that remain unsolved in cardiac quantification. We also added the performance of DMTRL both on the training set and the test set for comparison.

A. Performance on the validation dataset

As demonstrated in Table III, under the five-fold cross validation framework, all the submissions performed very well for all three LV indices and the cardiac phase on the training dataset, which were either clearly better or very close to the performance of the state-of-the-art model of DMTRL [41].

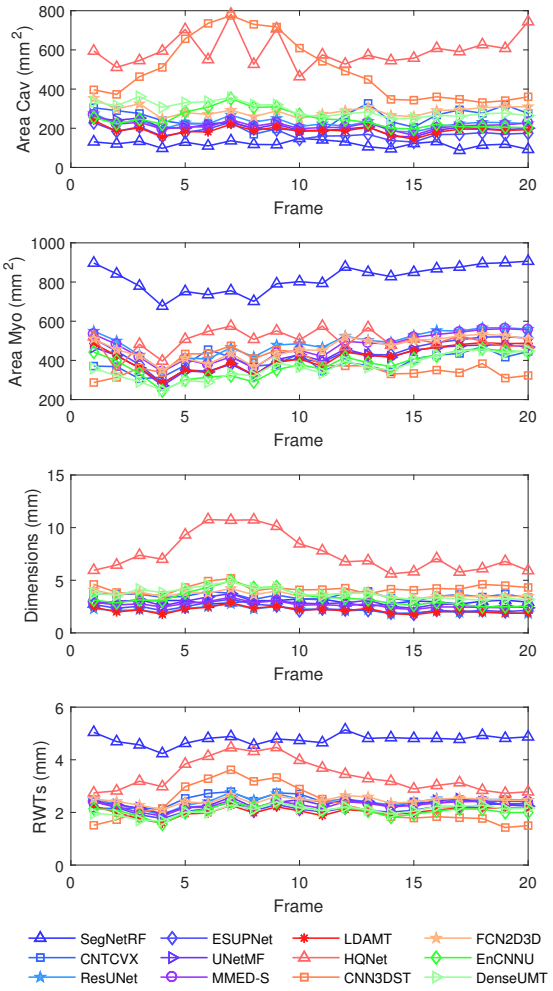

Fig. 4: Average frame-wise estimation errors of the LV indices for all submissions on the test dataset. The DR method LDAMT and the SG method ResUNet perform consistently well for all frames across the whole cardiac cycle and for all indices. (Similar colors represent methods from the same category.)

SG methods With the help of densely labeled myocardium masks for supervision during model training, the SG methods performed very well for all tasks. The best performance was achieved by ResUNet, with an average MAE of 62.3 mm^2, 0.79 mm, and 0.68 mm for areas, dimensions and RTWs, respectively, and a 6.72% ER for the cardiac phase. As a reference, the median and maximum values of these indices in the dataset are shown in Table II. This excellent performance could be attributed to the residual units in the Unet, which provided

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3064353, IEEE Journal of Biomedical and Health Informatics

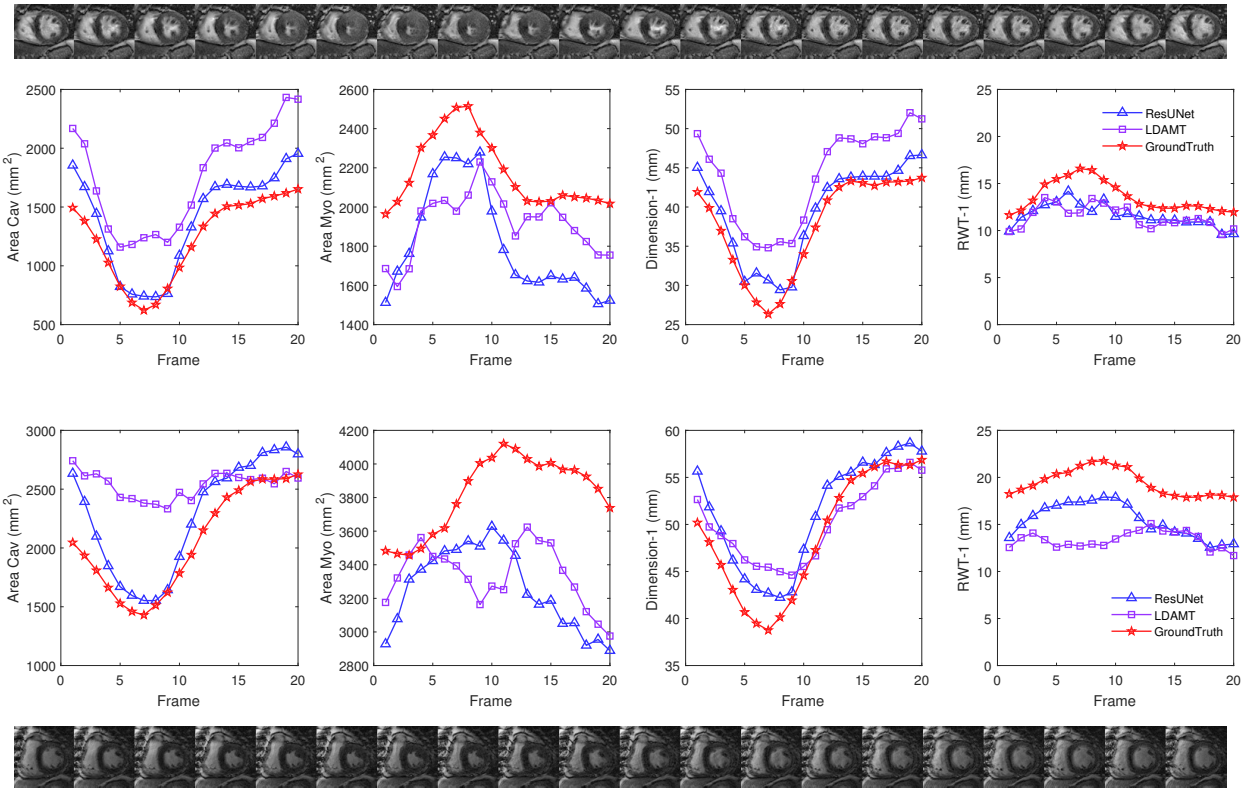

Fig. 5: Qualitative analysis of the SG method ResUNet and the DR method LADMT for a good case (top) and a bad case (bottom). Ground truth and prediction results of the whole cardiac sequences by the two methods are shown for A-cav, A-myo, dim1, RWT1, from left to right.

more diversity of information flow, and facilitated the network training [57]. Other methods that employ Unet and encoder-decoder architectures include ESUPNet, UNetMF, MMED-S, and SegNetRF. They also obtained similar performances. It is worth noting that CNTCVX was the only non-DNN-based segmentation method that used traditional image processing techniques, and it still obtained comparable performance to that of DMTRL on the training dataset.

DR methods With only the quantitative values of LV indices directly supervising the training procedure, the DR methods (LDAMT, MMED-R, HQNet, CNN3DST, and FCN2D3D), which are all based on DNNs, achieved inferior performances compared to the SG methods, while still being comparable to those of DMTRL. The best performance was achieved by LDAMT, with an average MAE of 156 mm^2, 2.03 mm, and 1.38 mm for area, dimensions and RTWs, and an 8.1% ER for the cardiac phase. It fused semantic and spatial information by aggregating layers with IDA and HDA, thus leading to a lower training error. MMED-R with the DR mode performed similarly to LDAMT and was better than the remaining DR methods.

Combined methods With ground truth values of both the segmentation mask and the cardiac indices during the network training, the combined methods performed slightly better than the DR methods. MMED-SR obtained the lowest MAE for dimensions (1.94 mm), EnCNNU for areas (124 mm^2), and DenseUMT for RWTs (1.37 mm). DenseUMT obtained the lowest error rate (7.8%) for cardiac phase prediction.

From these observations, we could draw that when validated on the training database: 1) the SG methods easily achieved better performance than the DR methods; 2) the DR ones require sophisticated network, e.g, LDA, to achieve good performance; 3) segmentation masks can help improve the performance of DR methods with more detailed pixel-level supervision, whereas the reverse is not necessarily true.

B. Performance on the test dataset

The generalization of all submissions were evaluated on the test dataset, and the results are shown in Table IV. For the three modes of MMED methods, only MMED-S is reported in the final test since it works best. As can be observed, the performances of the three types of methods obtain higher prediction error for four tasks on the test dataset than on the training dataset. Of the SG methods, none achieved the best performance for all tasks. ResUNet achieved the lowest MAE for areas (301 mm^2) and RWT (2.15 mm), whereas ESPUNet achieved the lowest MAE for dimensions (2.16 mm) and SegNetRF for the cardiac phase (9%). Of the DR methods, the best performance was achieved by LDAMT with a mean MAE of 301 mm^2, 2.15 mm, and 2.03 mm for areas, dimensions, and RWTs, respectively, and 9.5% for the cardiac phase. For the combined methods, EnCNNU achieved the lowest MAE for areas (311 mm^2) and dimensions (3.29 mm), while DenseUMT achieved the lowest MAE for RWTs (2.02

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3064353, IEEE Journal of Biomedical and Health Informatics

TABLE IV: Performance of all submissions on the test dataset (30 unseen subjects). MAE±std is shown for areas, dimensions and RWTs, and error rate is shown for cardiac phase. For each category, the best result is highlighted in boldface. The best DR method and the best SG methods perform very close to each other.

Methods	MAE of areas (mm^2)			MAE of Dimension (mm)				MAE of RWT (mm)							Phase
	A-cav	A-myo	mean	dim1	dim2	dim3	mean	IS	I	IL	AL	A	AS	mean	(%)
SG-based methods															
ResUNet	176 ±134	426 ±232	**301 ±183**	2.24 ±1.68	2.31 ±1.86	2.43 ±1.73	2.33 ±1.76	2.01 ±1.28	2.11 ±1.40	2.47 ±1.56	2.29 ±1.39	2.02 ±1.41	1.99 ±1.12	**2.15 ±1.36**	10
ESUPNet	225 ±168	490 ±233	357 ±201	2.38 ±1.87	2.21 ±1.92	1.87 ±1.43	**2.16 ±1.74**	2.58 ±1.36	2.7 ±1.45	2.77 ±1.57	2.27 ±1.52	2.03 ±1.63	2.14 ±1.26	2.41 ±1.47	12.7
UNetMF	204 ±140	413 ±200	309 ±170	2.7 ±3.35	2.7 ±3.47	3.14 ±3.52	2.85 ±3.45	1.84 ±1.24	2.53 ±1.46	2.52 ±1.57	2.43 ±1.60	2.54 ±1.59	2.07 ±1.45	2.31 ±1.48	12.33
MMED-S	214 ±139	469 ±216	341 ±178	2.68 ±1.73	2.61 ±1.68	2.72 ±1.70	2.67 ±1.70	2.16 ±1.26	2.57 ±1.39	2.62 ±1.44	2.1 ±1.25	2.07 ±1.25	2.13 ±1.12	2.28 ±1.29	10.33
SegNetRF	**119 ±95**	818 ±268	468 ±182	3.19 ±2.29	2.51 ±1.80	3.14 ±2.10	2.95 ±2.06	7.33 ±2.34	4.33 ±1.96	4.58 ±2.34	5.17 ±2.24	4.03 ±2.04	3.15 ±1.93	4.77 ±2.14	**9**
CNTCVX	249 ±169	**382 ±270**	320 ±220	2.7 ±1.99	4.65 ±3.07	3.21 ±2.06	3.51 ±2.37	2.78 ±1.82	1.88 ±1.44	1.78 ±1.42	2.01 ±1.68	2.26 ±1.67	3.74 ±1.93	2.41 ±1.66	11.2
DR-based methods															
LDAMT	**189 ±137**	413 ±194	**301 ±165**	2.21 ±1.56	2.11 ±1.64	2.12 ±1.63	**2.15 ±1.61**	2.05 ±1.13	2.09 ±1.33	2.25 ±1.32	1.89 ±1.21	1.9 ±1.20	2.08 ±1.16	**2.03 ±1.23**	**9.5**
HQNet	596 ±447	500 ±411	548 ±429	7.96 ±6.02	7.3 ±5.68	7.44 ±6.11	7.56 ±5.94	4.7 ±2.90	3.05 ±2.06	3.24 ±2.44	2.44 ±1.92	2.58 ±2.24	4.33 ±2.83	3.39 ±2.40	21.3
CNN3DST	494 ±350	**371 ±265**	432 ±308	4.04 ±3.15	4.43 ±3.41	4.19 ±3.13	4.22 ±3.23	2.58 ±2.04	2.12 ±1.80	2.5 ±2.01	2.04 ±1.82	1.93 ±1.70	2.46 ±1.91	2.27 ±1.88	19.5
FCN2D3D	288 ±189	463 ±256	375 ±223	3.81 ±2.35	3.57 ±2.31	3.56 ±2.40	3.65 ±2.35	2.58 ±1.53	2.47 ±1.52	2.63 ±1.79	2.18 ±1.51	2.27 ±1.44	2.57 ±1.50	2.45 ±1.55	12.83
Combined methods															
EnCNNU	**247 ±184**	377 ±228	**311 ±206**	3.61 ±2.58	3.3 ±2.39	2.97 ±2.33	**3.29 ±2.43**	2.14 ±1.42	2.35 ±1.56	2.46 ±1.60	1.99 ±1.36	1.67 ±1.22	1.76 ±1.25	**2.07 ±1.40**	15.67
DenseUMT	295 ±218	**363 ±261**	329 ±239	3.93 ±2.66	3.65 ±2.56	3.65 ±2.74	3.75 ±2.65	1.7 ±1.24	2.11 ±1.41	2.48 ±1.72	2.07 ±1.40	2.07 ±1.43	2.03 ±1.37	2.08 ±1.43	**10.5**
DMTRL	244 ±181	374 ±251	309 ±215	3.82 ±2.72	3.59 ±2.67	3.71 ±2.70	3.70 ±2.70	2.27 ±1.57	2.08 ±1.50	2.33 ±1.66	2.00 ±1.52	1.90 ±1.46	2.09 ±1.40	2.11 ±1.52	13.7

mm) and ER for the cardiac phase (10.5%). When compared to the state-of-the-art DMTRL, both ResUNet and LDAMT perform better for cavity area and dimensions and similar for RWTs.

According to the work of [20] in which inter-reader variability was studied with three human observers and 50 subjects, the inter-reader variability for LV cavity and myocardium contours ranged from 1.00 to 1.21 mm in terms of mean contour distance. Considering the fact that both the endocardium and epicardium contours must be accurate to obtain RWTs, and two points of the opposite direction on the endocardium define the dimension, the inter-reader variability for dimensions and RWTs are approximately 2.00 to 2.42 mm. The DR method LDAMT achieved the lowest estimation error for dimensions (2.15mm) and RWTs (2.03mm), which is actually comparable with the reported inter-reader variability in [20].

Fig. 4 shows the frame-wise estimation error of all submissions on the test dataset for cavity area, myocardium area, dimensions, and RWTs. As the plots show, SegNetRF achieved consistently the lowest estimation errors for the cavity area for all frames across the entire cardiac cycle, whereas it achieved the worst performance for the myocardium area. The network employed in SegNetRF tended to learn better endocardium borders then epicardium borders, which in turn led to poor estimation of RWTs. Most of the submissions performed generally well across the entire cardiac cycle, whereas HQNet and CNN3DST were prone to deliver higher estimation errors for frames close to ES, where the borders of myocardium can be easily affected by papillary muscles. LDAMT and ResUNet seemed to perform consistently well across the entire cardiac cycle and for all these indices.

We tested the significance of difference between different methods with paired student's t-tests for areas, dimensions, and RWTs, and the p-values are shown in Table V. We highlighted the p-values that are higher than 0.05. As can be drawn, most of the differences between two quantification methods are statistically significant. Considering the best SG method **ResUNet** and the best DR method **LDAMT**, the difference in performance is insignificant for areas and significant for dimensions and RWTs. The prediction results of these two methods for two cases in the test dataset are shown in Fig. 5. For the good case who has clear boundary between the myocardium and the surrounding background, both methods can capture the variations well across the whole cardiac cycle. ResUNet estimates the A-cav and dimensions better than LADMT, while the latter performs better for A-myo. For the bad case whose lateral boundaries are nearly invisible, both methods fails for A-myo, and RWT, and ResUNet performs better for A-cav, due the high contrast between the cavity and the myocardium.

A comparison between the validation performance on the training dataset and the test performance indicates how these submissions generalize to new data. The results in Tables III and IV reveal that lower validation errors on the training dataset did not necessarily lead to better generalization. For example, ESUPNet and MMED-S achieved lower generalization than LDAMT for all the four tasks, despite their better performance on the training dataset. CNN3DST obtained a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3064353, IEEE Journal of Biomedical and Health Informatics

high ER for the cardiac phase prediction but achieved the lowest ER for the training dataset.

C. Discussions on SG and DR methods

SG vs. DR methods The aforementioned results demonstrated that with the common goal of cardiac indices quantification, both SG and DR methods have great potential to obtain accurate estimation. In clinical practice, automatic cardiac quantification methods have the potential to help alleviate the tedious workload of manually contouring and measurement of important cardiac indices from plenty of imaging data. On one hand, analysis from these imaging data, such as systolic function evaluation, cardiac event monitoring, hypertrophy categorization, remodeling indication, regional dysfunction indication, can be completed with a greatly improved efficiency with the results of automatic quantification methods. Therefore, the whole clinical workflow, including diagnosis, progress monitoring, decision making and treatment evaluation, can all be advanced. On the other hand, automatic methods can reduce the observer variation, thus improve the clinical value of those quantification results. Besides, reliable and interpretable quantification results allow experts' visual inspecting and make it easy to explain, therefore user's understanding and trust can be easily built, which is critical for practical application of automatic methods.

However, both types of methods possess advantages and disadvantages. As mentioned in [49], SG methods allow straight visual assessment of a network's outputs and thus facilitate the identification of failed cases, and guide proposing of new algorithms for improvement. The segmentation results may add values for other applications including image-guided cardiac interventions [52]. The segmentation and computation steps make the whole quantification procedure less of a black box, and easy to understand and explain. Such a outcome explanation is important for users to build trust on the computer-assisted system [64]. By contrast, DR methods estimate directly the final quantification values from the appearance of cardiac images, which is more like a black box, regardless of the image feature or network architecture that is employed. In addition, DR methods apparently cannot provide visual inspection.

However, DR methods for cardiac quantification in themselves can be viewed as great success given the fact that they circumvent the requirement of numerous densely labeled images. In clinical practice, cardiac physicians produce reports by first describing the appearance of the myocardium and its motion abnormality locally and globally, and then recording the obtained measurements. These measurements in the reports can be readily employed for DR methods to learn the quantification model, therefore making it easy to include other slices and conduct large-scale and multi-center studies. Besides, the additional measuring step for SG methods to obtain the final quantification results is not required in DR methods.

Combination of SG and DR It is worth noting the effectiveness of the combination strategy, which was also encouraged during the challenge to maximally leverage the supervisory information from both the binary segmentation mask and quantitative ground truth values. Of the three combined methods, DenseUMT and EnCNNU achieved improved performance when the two tasks were combined. For MMED, the combined method of MMED-SR obtained better performance only over its DR counterpart MMED-R, and inferior performance to its SG counterpart MMED-S. This inferior performance of MMED-SR may have been because the two branches share the same Unet for feature extraction, which was more suitable for and more frequently updated by the segmentation task. This resulted in features with greater discrimination capability for the segmentation and less expressiveness for the regression. By contrast, the two tasks did not share parameters in DenseUMT and EnCNNU. The segmentation task provided a warm start for the CNN embedding module to learn regression features in DenseUMT. In EnCNNU, the segmentation and regression were implemented in two separate networks, with one achieving better estimation for areas and dimensions and the other for RTWs. The two networks complemented each other well and delivered improved accuracy when combined by a gradient-boosting-based ensemble learning algorithm.

Given these discussions, combining the advantages of both SG and DR methods has a great promise for the task of LV quantification. A method with visualization ability learned from a few densely labeled samples and quantification ability learned from large-scale records of measurement will be a flexible way for interpretable and accurate clinical application.

VI. Conclusion

Given the clinical significance of the task, the LV quantification challenge attracted a wide global interest to achieve advancements in the area of LV quantification. With the leading techniques in machine learning, the difficulties related to robust image representation, and the temporal deformation of myocardium for cardiac images were effectively alleviated, and excellent LV quantification performances were obtained by the submissions. Evaluation of these submissions demonstrated that 1) both SG and DR methods can achieve good and comparable performance, whereas DR methods can relieve the dependency on densely labeled binary masks; 2) to obtain stable and accurate prediction, more attentions must be given to the design of network architecture in DR methods than to that in SG methods to extract robust representations of anatomic structures. The dataset for this challenge will remain open to the community to encourage more advancements in cardiac quantification regrading this aspect.

Despite the good performance achieved in this challenge, problems remain to be solved for real application of cardiac quantification in routine practice. First, only one slice of the mid-cavity was included in this challenge, which means that only 2D evaluation results were obtained. Full stack slices must be included to obtain accurate 3D evaluations. In this case, inter-slice dependency should be explored, for example, by means of recurrent neural network, sequential learning [65], or label propagation. Anatomical priors of the basal and apical slices can also be introduced into the model to further improve the performance. Second, in addition to the four tasks considered in the LVQuan Challenge, other clinical indices such as myocardium strain and regional motion are also

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3064353, IEEE Journal of Biomedical and Health Informatics

TABLE V: p-values of paired student's t-test are demonstrated to test the significance of performance difference for different methods. p-values higher than 0.05 are highlighted in bold face.

	ESUPNet	UNetMF	MMED-S	SegNetRF	CNTCVX	LDAMT	HQNet	CNN3DST	FCN2D3D	EnCNNU	DenseUMT
Areas											
ResUNet	<0.001	**0.41**	<0.001	<0.001	0.042	**0.98**	<0.001	<0.001	<0.001	**0.25**	0.0033
ESUPNet		<0.001	**0.10**	<0.001	<0.001	<0.001	<0.001	<0.001	**0.07**	<0.001	0.0062
UNetMF			<0.001	<0.001	**0.20**	**0.38**	<0.001	<0.001	<0.001	**0.74**	0.028
MMED-S				<0.001	0.03	<0.001	<0.001	<0.001	<0.001	<0.001	**0.22**
SegNetRF					<0.001	<0.001	<0.001	0.0013	<0.001	<0.001	<0.001
CNTCVX						0.034	<0.001	<0.001	<0.001	**0.33**	**0.38**
LDAMT							<0.001	<0.001	<0.001	**0.22**	0.0023
HQNet								<0.001	<0.001	<0.001	<0.001
CNN3DST									<0.001	<0.001	<0.001
FCN2D3D										<0.001	<0.001
EnCNNU											**0.057**
Dimensions											
ResUNet	0.041	<0.001	<0.001	<0.001	<0.001	0.031	<0.001	<0.001	<0.001	<0.001	<0.001
ESUPNet		<0.001	<0.001	<0.001	<0.001	**0.97**	<0.001	<0.001	<0.001	<0.001	<0.001
UNetMF			0.034	**0.25**	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
MMED-S				<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
SegNetRF					<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
CNTCVX						<0.001	<0.001	<0.001	**0.21**	0.036	**0.05**
LDAMT							<0.001	<0.001	<0.001	<0.001	<0.001
HQNet								<0.001	<0.001	<0.001	<0.001
CNN3DST									<0.001	<0.001	0.0011
FCN2D3D										0.00021	**0.44**
EnCNNU											<0.001
RWTs											
ResUNet	<0.001	<0.001	0.0092	<0.001	<0.001	0.014	<0.001	**0.087**	<0.001	**0.09**	**0.17**
ESUPNet		**0.05**	0.0063	<0.001	**0.9**	<0.001	<0.001	0.047	**0.55**	<0.001	<0.001
UNetMF			**0.41**	<0.001	**0.062**	<0.001	<0.001	**0.52**	0.014	<0.001	<0.001
MMED-S				<0.001	0.0083	<0.001	<0.001	**0.94**	0.0014	<0.001	<0.001
SegNetRF					<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
CNTCVX						<0.001	<0.001	**0.055**	**0.49**	<0.001	<0.001
LDAMT							<0.001	<0.001	<0.001	**0.55**	**0.38**
HQNet								<0.001	<0.001	<0.001	<0.001
CNN3DST									0.018	0.0038	0.0077
FCN2D3D										<0.001	<0.001
EnCNNU											**0.78**

major indicators of cardiac function and should be included to deliver a comprehensive evaluation of cardiac function. Third, quantification of these cardiac indices is not the final task. Even though not directly, the metrics considered in this challenge are necessary for the calculation of LV EF, cavity volume, LV mass, and evaluation of regional wall thickening and myocardium remodeling. The relation between them and cardiac function must also be analyzed, so that the automatic quantification can be applied to all stages of cardiac disease, including diagnosis, evaluation, monitoring and prognosis to improve the entire work flow of cardiac disease treatment.

ACKNOWLEDGMENTS

The paper is partially supported by the Natural Science Foundation of China under Grants 61801296. The work of Eric Kerfoot was supported by an EPSRC programme Grant (EP/P001009/1) and the Wellcome EPSRC Centre for Medical Engineering at the School of Biomedical Engineering and Imaging Sciences, Kings College London (WT 203148/Z/16/Z). The work of Angélica Atehortúa was supported by Colciencias-Colombia, Grant No. 647 (2015 call for National PhD studies) and Université de Rennes 1. The work of Alejandro Debus was supported by the Santa Fe Science, Technology and Innovation Agency (AS ACTEI), Government of the Province of Santa Fe, through Project AC-00010-18, Resolution N 117/14.

REFERENCES

[1] W. H. Organization *et al.*, *Global status report on noncommunicable diseases 2014*. World Health Organization, 2014, no. WHO/NMH/NVI/15.1.

[2] A. Suinesiaputra, B. R. Cowan, A. O. Al-Agamy, M. A. Elattar, N. Ayache, A. S. Fahmy, A. M. Khalifa, P. Medrano-Gracia, M.-P. Jolly, A. H. Kadish, D. C. Lee, J. Margeta, S. K. Warfield, and A. A. Young, "A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images," *Medical Image Analysis*, vol. 18, no. 1, pp. 50 – 62, 2014.

[3] T. D. Karamitsos, J. M. Francis, S. Myerson, J. B. Selvanayagam, and S. Neubauer, "The role of cardiovascular magnetic resonance imaging in heart failure," *Journal of the American College of Cardiology*, vol. 54, no. 15, pp. 1407–1424, 2009.

[4] J. Schulz-Menger, D. A. Bluemke, J. Bremerich, S. D. Flamm, M. A. Fogel, M. G. Friedrich, R. J. Kim, F. von Knobelsdorff-Brenkenhoff, C. M. Kramer, D. J. Pennell *et al.*, "Standardized image interpretation and post processing in cardiovascular magnetic resonance: Society for cardiovascular magnetic resonance (SCMR) board of trustees task force on standardized post processing," *Journal of Cardiovascular Magnetic Resonance*, vol. 15, no. 1, p. 35, 2013.

[5] R. M. Lang, M. Bierig, R. B. Devereux, F. A. Flachskampf, E. Foster, P. A. Pellikka, M. H. Picard, M. J. Roman, J. Seward, J. Shanewise *et al.*, "Recommendations for chamber quantification," *European journal of echocardiography*, vol. 7, no. 2, pp. 79–108, 2006.

[6] P. Peng, K. Lekadir, A. Gooya, L. Shao, S. E. Petersen, and A. F. Frangi, "A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 29, no. 2, pp. 155–195, 2016.

[7] V. O. Puntmann, R. Gebker, S. Duckett, J. Mirelis, B. Schnackenburg, M. Graefe, R. Razavi, E. Fleck, and E. Nagel, "Left ventricular chamber dimensions and wall thickness by cardiovascular magnetic

resonance: comparison with transthoracic echocardiography," *European Heart Journal–Cardiovascular Imaging*, vol. 14, no. 3, pp. 240–246, 2013.

[8] N. Kawel, E. B. Turkbey, J. J. Carr, J. Eng, A. S. Gomes, W. G. Hundley, C. Johnson, S. C. Masri, M. R. Prince, R. J. van der Geest *et al.*, "Normal left ventricular myocardial thickness for middle-aged and older subjects with steady-state free precession cardiac magnetic resonance: the multi-ethnic study of atherosclerosis," *Circulation: Cardiovascular Imaging*, vol. 5, no. 4, pp. 500–508, 2012.

[9] C. Petitjean and J.-N. Dacher, "A review of segmentation methods in short axis cardiac MR images," *Medical Image Analysis*, vol. 15, no. 2, pp. 169–184, 2011.

[10] A. Gupta, L. Von Kurowski, A. Singh, D. Geiger, C.-C. Liang, M.-Y. Chiu, L. Adler, M. Haacke, and D. Wilson, "Cardiac MR image segmentation using deformable models," in *Computers in Cardiology 1993, Proceedings.* IEEE, 1993, pp. 747–750.

[11] E. Nachtomy, R. Cooperstein, M. Vaturi, E. Bosak, Z. Vered, and S. Akselrod, "Automatic assessment of cardiac function from short-axis MRI: procedure and clinical evaluation," *Magnetic resonance imaging*, vol. 16, no. 4, pp. 365–376, 1998.

[12] I. B. Ayed, H.-M. Chen, K. Punithakumar, I. Ross, and S. Li, "Max-flow segmentation of the left ventricle by recovering subject-specific distributions via a bound of the bhattacharyya measure," *Medical image analysis*, vol. 16, no. 1, pp. 87–100, 2012.

[13] Y. Wu, Y. Wang, and Y. Jia, "Segmentation of the left ventricle in cardiac cine MRI using a shape-constrained snake model," *Computer Vision and Image Understanding*, vol. 117, no. 9, pp. 990–1003, 2013.

[14] A. Pednekar, U. Kurkure, R. Muthupillai, S. Flamm, and I. A. Kakadiaris, "Automated left ventricular segmentation in cardiac MRI," *IEEE TBME*, vol. 53, no. 7, pp. 1425–1428, 2006.

[15] J. Lötjönen, S. Kivistö, J. Koikkalainen, D. Smutek, and K. Lauerma, "Statistical shape model of atria, ventricles and epicardium from short- and long-axis MR images," *Medical image analysis*, vol. 8, no. 3, pp. 371–386, 2004.

[16] M. Avendi, A. Kheradvar, and H. Jafarkhani, "A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI," *Medical image analysis*, vol. 30, pp. 108–119, 2016.

[17] T. A. Ngo and G. Carneiro, "Left ventricle segmentation from cardiac MRI combining level set methods with deep belief networks," in *ICIP*. IEEE, 2013, pp. 695–699.

[18] T. A. Ngo, Z. Lu, and G. Carneiro, "Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance," *Medical image analysis*, vol. 35, pp. 159–171, 2017.

[19] P. V. Tran, "A fully convolutional neural network for cardiac segmentation in short-axis MRI," *arXiv:1604.00494*, 2016.

[20] W. Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vaillant, A. M. Lee, N. Aung, E. Lukaschuk, M. M. Sanghvi *et al.*, "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks," *Journal of Cardiovascular Magnetic Resonance*, vol. 20, no. 1, p. 65, 2018.

[21] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE TPAMI*, vol. 39, no. 4, pp. 640–651, 2017.

[22] J. Patravali, S. Jain, and S. Chilamkurthy, "2D-3D fully convolutional neural networks for cardiac MR segmentation," *STACOM*, 2017.

[23] C. Zotti, Z. Luo, A. Lalande, O. Humbert, and P.-M. Jodoin, "Novel deep convolution neural network applied to MRI cardiac segmentation," *preprint arXiv:1705.08943*, 2017.

[24] C. Zotti, Z. Luo, A. Lalande, and P.-M. Jodoin, "Convolutional neural network with shape prior applied to cardiac MRI segmentation," *IEEE JBHI*, 2019.

[25] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Isgum, "Automatic segmentation and disease classification using cardiac cine MR images," *STACOM*, 2017.

[26] A. Mortazi, J. Burt, and U. Bagci, "Multi-planar deep segmentation networks for cardiac substructures from MRI and CT," *STACOM*, 2017.

[27] L. K. Tan, Y. M. Liew, E. Lim, and R. A. McLaughlin, "Convolutional neural network regression for short-axis left ventricle segmentation in cardiac cine MR sequences," *Medical Image Analysis*, vol. 39, pp. 78–86, 2017.

[28] Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin, and P.-A. Heng, "3D deeply supervised network for automated segmentation of volumetric medical images," *Medical Image Analysis*, vol. 41, pp. 40–54, 2017.

[29] S. Dangi, C. A. Linte, and Z. Yaniv, "A distance map regularized CNN for cardiac cine MR image segmentation," *Medical physics*, vol. 46, no. 12, pp. 5637–5651, 2019.

[30] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert, "Deep learning for cardiac image segmentation: A review," *Frontiers in Cardiovascular Medicine*, vol. 7, p. 25, 2020.

[31] M. Afshin, I. B. Ayed, A. Islam, A. Goela, T. M. Peters, and S. Li, "Global assessment of cardiac function using image statistics in MRI," in *MICCAI*. Springer, 2012, pp. 535–543.

[32] M. Afshin, I. Ben Ayed, K. Punithakumar, M. Law, A. Islam, A. Goela, T. M. Peters, and S. Li, "Regional assessment of cardiac left ventricular myocardial function via MRI statistical features," *IEEE TMI*, vol. 33, no. 2, pp. 481–494, 2014.

[33] Z. Wang, M. Ben Salah, B. Gu, A. Islam, A. Goela, and S. Li, "Direct estimation of cardiac biventricular volumes with an adapted bayesian formulation," *IEEE TBE*, vol. 61, no. 4, pp. 1251–1260, 2014.

[34] X. Zhen, Z. Wang, A. Islam, M. Bhaduri, I. Chan, and S. Li, "Direct estimation of cardiac bi-ventricular volumes with regression forests," in *MICCAI*. Springer, 2014, pp. 586–593.

[35] ——, "Multi-scale deep networks and regression forests for direct bi-ventricular volume estimation," *Medical image analysis*, vol. 30, pp. 120–129, 2016.

[36] X. Zhen, A. Islam, M. Bhaduri, I. Chan, and S. Li, "Direct and simultaneous four-chamber volume estimation by multi-output regression," in *MICCAI*. Springer, 2015, pp. 669–676.

[37] A. Kabani and M. R. El-Sakka, "Estimating ejection fraction and left ventricle volume using deep convolutional networks," in *International Conference Image Analysis and Recognition*, 2016, pp. 678–686.

[38] W. Xue, A. Islam, M. Bhaduri, and S. Li, "Direct multitype cardiac indices estimation via joint representation and regression learning." *IEEE TMI*, vol. 36, no. 10, p. 2057, 2017.

[39] W. Xue, I. B. Nachum, S. Pandey, J. Warrington, S. Leung, and S. Li, "Direct estimation of regional wall thicknesses via residual recurrent neural network," in *IPMI*. Springer, 2017, pp. 505–516.

[40] W. Xue, A. Lum, A. Mercado, M. Landis, J. Warrington, and S. Li, "Full quantification of left ventricle via deep multitask learning network respecting intra- and inter-task relatedness," in *MICCAI*. Springer, 2017.

[41] W. Xue, G. Brahm, S. Pandey, S. Leung, and S. Li, "Full left ventricle quantification via deep multitask relationships learning," *Medical Image Analysis*, vol. 43, pp. 54 – 65, 2018.

[42] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M. Rohé, X. Pennec, M. Sermesant, F. Isensee, P. Jäger, K. H. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, I. Išgum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, and P. Jodoin, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE TMI*, vol. 37, no. 11, pp. 2514–2525, Nov 2018.

[43] S. E. Petersen, P. M. Matthews, J. M. Francis, M. D. Robson, F. Zemrak, R. Boubertakh, A. A. Young, S. Hudson, P. Weale, S. Garratt *et al.*, "UK biobank's cardiovascular magnetic resonance protocol," *Journal of cardiovascular magnetic resonance*, vol. 18, no. 1, p. 8, 2015.

[44] M. D. Cerqueira, N. J. Weissman, V. Dilsizian, A. K. Jacobs, S. Kaul, W. K. Laskey, D. J. Pennell, J. A. Rumberger, T. Ryan, M. S. Verani *et al.*, "Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart a statement for healthcare professionals from the cardiac imaging committee of the council on clinical cardiology of the american heart association," *Circulation*, vol. 105, no. 4, pp. 539–542, 2002.

[45] C. W. Yancy, M. Jessup, B. Bozkurt, J. Butler, D. E. Casey, M. H. Drazner, G. C. Fonarow, S. A. Geraci, T. Horwich, J. L. Januzzi *et al.*, "2013 ACCF/AHA guideline for the management of heart failure: a report of the american college of cardiology foundation/american heart association task force on practice guidelines," *Journal of the American College of Cardiology*, vol. 62, no. 16, pp. e147–e239, 2013.

[46] A. Suinesiaputra, D. A. Bluemke, B. R. Cowan, M. G. Friedrich, C. M. Kramer, R. Kwong, S. Plein, J. Schulz-Menger, J. J. Westenberg, A. A. Young *et al.*, "Quantification of LV function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours," *Journal of Cardiovascular Magnetic Resonance*, vol. 17, no. 1, p. 63, 2015.

[47] V. G. Buller, R. J. Van Der Geest, M. D. Kool, E. E. Van Der Wall, A. De Roos, and J. H. Reiber, "Assessment of regional left ventricular wall parameters from short axis magnetic resonance imaging using

a three-dimensional extension to the improved centerline method," *Investigative radiology*, vol. 32, no. 9, pp. 529–539, 1997.

[48] M. Pop, M. Sermesant, J. Zhao, S. Li, K. McLeod, A. Young, K. Rhode, and T. Mansi, *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges: 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018*. Springer, 2019, vol. 11395.

[49] E. Kerfoot, J. Clough, I. Oksuz, J. Lee, A. P. King, and J. A. Schnabel, "Left-ventricle quantification using residual U-net," in *STACOM*, 2018, pp. 371–380.

[50] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[51] W. Yan, Y. Wang, S. Chen, R. J. van der Geest, and Q. Tao, "ESU-P-Net: Cascading network for full quantification of left ventricle from cine mri," in *STACOM*, 2018, pp. 421–428.

[52] F. Guo, M. Ng, and G. Wright, "Cardiac MRI left ventricle segmentation and quantification: A framework combining U-net and continuous max-flow," in *STACOM*, 2018, pp. 450–458.

[53] A. Atehortúa, M. Garreau, D. Romo-Bucheli, and E. Romero, "Automatic left ventricle quantification in cardiac MRI via hierarchical refinement of high-level features by a salient perceptual grouping model," in *STACOM*, 2018, pp. 439–449.

[54] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE TPAMI*, vol. 39, no. 12, pp. 2481–2495, Dec 2017.

[55] E. Grinias and G. Tziritas, "Convexity and connectivity principles applied for left ventricle segmentation and quantification," in *STACOM*, 2018, pp. 389–401.

[56] J. Li and Z. Hu, "Left ventricle full quantification using deep layer aggregation based multitask relationship," in *STACOM*, 2018, pp. 381–388.

[57] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *IEEE CVPR*, June 2018.

[58] G. Yang, T. Hua, C. Lu, T. Pan, X. Yang, L. Hu, J. Wu, X. Zhu, and H. Shu, "Left ventricle full quantification via hierarchical quantification network," in *STACOM*, 2018, pp. 429–438.

[59] A. Debus and E. Ferrante, "Left ventricle quantification through spatio-temporal CNNs," in *STACOM*, 2018, pp. 466–475.

[60] Y. Jang, S. Kim, H. Shim, and H.-J. Chang, "Full quantification of left ventricle using deep multitask network with combination of 2D and 3D convolution on 2D + t cine MRI," in *STACOM*, 2018, pp. 476–483.

[61] H. Xu, J. E. Schneider, and V. Grau, "Calculation of anatomical and functional metrics using deep learning in cardiac MRI: Comparison between direct and segmentation-based estimation," in *STACOM*, 2018, pp. 402–411.

[62] J. Liu, X. Li, H. Ren, and Q. Li, "Multi-estimator full left ventricle quantification through ensemble learning," in *STACOM*, 2018, pp. 459–465.

[63] L. Liu, J. Ma, J. Wang, and J. Xiao, "Automated full quantification of left ventricle with deep neural networks," in *STACOM*, 2018, pp. 412–420.

[64] J. Stoyanovich, J. V. V. Bavel, and T. V. West, "The imperative of interpretable machines," *Nature Machine Intelligence*, vol. 2, pp. 197–199, 2020.

[65] Y.-L. Lu, K. A. Connelly, A. J. Dick, G. A. Wright, and P. E. Radau, "Automatic functional analysis of left ventricle in cardiac cine MRI," *Quantitative imaging in medicine and surgery*, vol. 3, no. 4, p. 200, 2013.