Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography

Sarah Leclerc, Erik Smistad, João Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, Carole Lartizien, Jan D'hooge, Lasse Lovstakken, and Olivier Bernard

***Abstract*— Delineation of the cardiac structures from 2D echocardiographic images is a common clinical task to establish a diagnosis. Over the past decades, the automation of this task has been the subject of intense research. In this paper, we evaluate how far the state-of-the-art encoder-decoder deep convolutional neural network methods can go at assessing 2D echocardiographic images, i.e. segmenting cardiac structures and estimating clinical indices, on a dataset, especially, designed to answer this objective. We, therefore, introduce the cardiac acquisitions for multi-structure ultrasound segmentation dataset, the largest publicly-available and fully-annotated dataset for the purpose of echocardiographic assessment. The dataset contains two and four-chamber acquisitions from 500 patients with reference measurements from one cardiologist on the full dataset and from three cardiologists on a fold of 50 patients. Results show that encoder-decoder-based architectures outperform state-of-the-art non-deep learning methods and faithfully reproduce the expert analysis for the end-diastolic and end-systolic left ventricular volumes, with a mean correlation of 0.95 and an absolute mean error of 9.5 ml. Concerning the ejection fraction of the left ventricle, results are more contrasted with a mean correlation coefficient of 0.80 and an absolute mean error of 5.6%. Although these results are below the inter-observer scores, they remain slightly worse than the intra-observer's ones. Based on this observation, areas for improvement are defined, which open the door for accurate and fully-automatic analysis of 2D echocardiographic images.**

***Index Terms*— Cardiac segmentation and diagnosis, deep learning, ultrasound, left ventricle, myocardium, left atrium.**

Manuscript received December 19, 2018; revised February 14, 2019; accepted February 14, 2019. Date of publication February 22, 2019; date of current version August 30, 2019. This work was supported in part by the framework of the LABEX PRIMES (ANR-11-LABX-0063) of the Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR), and in part by the Centre for Innovative Ultrasound Solutions through the Norwegian Research Council under Project 237887. *(Corresponding author: Sarah Leclerc.)*

S. Leclerc, F. Cervenansky, T. Grenier, C. Lartizien, and O. Bernard are with the University of Lyon, CREATIS, CNRS UMR5220, Inserm U1044, INSA-Lyon, University of Lyon 1, 69100 Villeurbanne, France (e-mail: sarah.leclerc@gmx.fr; olivier.bernard@creatis.insa-lyon.fr).

E. Smistad, A. Ostvik, and L. Lovstakken are with the Center of Innovative Ultrasound Solutions, Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, 7491 Trondheim, Norway.

J. Pedrosa and J. D'hooge are with the Department of Cardiovascular Sciences, KU Leuven, 3000 Leuven, Belgium.

F. Espinosa is with the Cardiovascular Department, Centre Hospitalier Universitaire de Saint-Etienne, 42270 Saint-Etienne, France.

T. Espeland and E. A. Rye Berg are with the Center of Innovative Ultrasound Solutions and the Clinic of Cardiology, St. Olavs Hospital, 7030 Trondheim, Norway.

P.-M. Jodoin is with the Computer Science Department, University of Sherbrooke, Sherbrooke, QC J1K2R1, Canada.

This article has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2019.2900516

I. Introduction

ANALYSIS of 2D echocardiographic images plays a crucial role in clinical routine to measure the cardiac morphology and function and to reach a diagnosis. Such analysis is based on the interpretation of clinical indices which are extracted from low-level image processing such as segmentation and tracking. For instance, the extraction of the ejection fraction (EF) of the left ventricle (LV) requires accurate delineation of the left ventricular endocardium in both end diastole (ED) and end systole (ES). In clinical routine, semi-automatic or manual annotation is still daily work due to the lack of accuracy and reproducibility of fully-automatic cardiac segmentation methods. This leads to time consuming tasks prone to intra- and inter-observer variability [1]. The inherent difficulties for segmenting echocardiographic images are well documented: *i)* poor contrast between the myocardium and the blood pool; *ii)* brightness inhomogeneities; *iii)* variation in the speckle pattern along the myocardium due to the orientation of the cardiac probe with respect to tissue; *iv)* presence of trabeculae and papillary muscles with intensities similar to the myocardium; *v)* significant tissue echogenicity variability within the population; *vi)* shape, intensity and motion variability of the heart structures across patients and pathologies.

The lack of large and publicly-available dataset has prevented a thorough evaluation of the potential of deep learning methods to estimate clinical indices, while these techniques are actively applied with great success for other modalities [2]. Indeed, while the number of medical imaging challenges comparing deep learning methods has exploded this last decade, only one focused on cardiac ultrasound image segmentation [3]. Unfortunately, since the challenge was held

in 2014, none of the participant used convolutional neural networks (CNNs) because these methods had not yet gained popularity in medical imaging. The lack of well-annotated echocardiographic datasets can be explained by the difficulty of exporting data from clinical ultrasound equipments and getting a large amount of images carefully annotated by cardiologists due to the very nature of echocardiography as mentioned above. In this context, the purpose of this paper is to provide answers to the following four questions:

1) How well do CNNs perform compared to non-deep learning state-of-the-art techniques ?
2) How many patients are needed to train a CNN to get highly accurate results in 2D echocardiographic image segmentation ?
3) How accurate can the volumes and ejection fraction be estimated from the segmentation of CNNs compared to the inter/intra-expert variability ?
4) What improvement can be brought by sophisticated architectures compared to simpler CNN models for 2D echocardiographic segmentation ?

For that purpose, we present a new public dataset called CAMUS (Cardiac Acquisitions for Multi-structure Ultrasound Segmentation). It contains 2D echocardiographic sequences with two and four-chamber views of 500 patients that were acquired with the same equipment in the same medical center. The size of this dataset and its tight connection to every-day clinical issues give the possibility to train deep learning methods to automatically analyze echocardiographic data. In addition, CAMUS includes manual expert annotations for the left ventricle endocardium (LV_{Endo}), the myocardium (epicardium contour more specifically, named LV_{Epi}) and the left atrium (LA).

II. Previous Work

A. Previous Cardiac Ultrasound Datasets

To date, only one echocardiographic dataset has been broadly validated. This dataset was released in conjunction with the Challenge on Endocardial Three-dimensional Ultrasound Segmentation (CETUS) which took place during the MICCAI 2014 conference.[1] The CETUS dataset is composed of 45 3D echocardiographic sequences (15 for training, 30 for testing) equally distributed among three different subgroups: healthy subjects, patients with previous myocardial infarction examined at least 3 months after the event and patients with dilated cardiomyopathy. The data is provided with two reference meshes of the LV_{Endo} per patient (one at ED and one at ES), each reference corresponding to the mean shape computed from the annotations of three different experienced cardiologists. Five fully-automatic (deformable models, Hough random forest, Kalman filter, active appearance model) and four semi-automatic methods (graph-cut method, structured random forest, multi-atlas and level-set approaches) were evaluated through this challenge. No challenger implemented a deep neural network. The outcome of the challenge revealed that the overall best scores were obtained by the B-spline explicit active surface, a fully-automatic method proposed by Barbosa *et al.* [4]. This method was later on improved by Pedrosa *et al.* [5] thanks to the integration of a shape prior derived from a conventional principal component analysis scheme. By doing so, the authors obtained the following scores for the segmentation of the 3D LV_{Endo}: *i)* average Dice values of 0.909 (ED) and 0.875 (ES); *ii)* average Hausdorff distances of 6.3 mm (ED) and 6.9 mm (ES) and *iii)* average mean absolute distances of 1.8 mm (ED) and 2.0 mm (ES).

[1] https://www.creatis.insa-lyon.fr/Challenge/CETUS/

B. Non-Deep Learning Methods

Several surveys of echocardiographic segmentation methods have been proposed, both in 2D [6], [7] and 3D [3], [8]. Most of the reported methods focused on the segmentation of the LV_{Endo} border. Among those reviews, only the one by Bernard *et al.* [3] published in 2016 benchmarked different techniques on the same dataset, leading to a fair comparison. In this study, the authors listed the results obtained by nine different techniques. The reported methods can be divided in two main categories: those with a weak prior and those with a strong prior. The first group involves weak assumptions such as spatial, intensity, motion or anatomical information. It includes image-based techniques (multi-scale quadrature filter) [9], a motion-based method (Kalman filter) [10], deformable models (BEAS, level-set) [4], [9] and a graph-based approach (graph-cut) [11]. The second group uses approaches with strong priors including a shape-prior (Hough forest) [12], an active appearance model [13], an atlas-based method [14] and a machine learning algorithm (random forest) [12], [15], [16], each requiring a manually-annotated training set.

C. Deep Learning Methods

Deep-learning methods have been successfully applied to the segmentation of the LV_{Endo} in echocardiography. In 2012, Carneiro *et al.* developed a two-stage deep learning method for the segmentation of the LV_{Endo} for 2D echocardiographic images restricted to four-chamber view acquisitions [7]. Based on a maximum *a posteriori* framework, the authors formulated the LV segmentation problem according to two successive steps: *i)* the automatic selection of several regions in the tested image where the LV_{Endo} is fully present; *ii)* the automatic extraction of the LV_{Endo} contour from the previously selected regions. These two steps involved a deep belief network. Their method was trained on 400 images from 12 different patient sequences with various pathologies and tested on 50 images from 2 healthy subject sequences. They obtained an average Hausdorff distance of $\sim$18 mm and an average mean absolute distance of $\sim$8 mm for the LV_{Endo}.

In 2017, Smistad *et al.* [17] showed that the U-Net CNN method [18] could be trained to successfully segment the left ventricle in 2D ultrasound images. However, due to lack of training data, the network was trained with the output of a state-of-the-art deformable model segmentation method [10]. On a manually segmented test set, the results showed that the network and the deformable model obtained the same accuracy with a Dice score of 0.87.

Recently, Oktay *et al.* [19] used CNNs to segment the 3D LV_{Endo} structure using an approach named anatomically constrained neural network (ACNN). The core of their neural network is based on an architecture similar to the 3D U-Net [20], whose segmentation output is constrained to fit a non-linear compact representation of the underlying anatomy derived from an auto-encoder network. The performance of their method was assessed on the CETUS dataset. They obtained the following scores for the segmentation of the 3D LV_{Endo} structure: *i)* average Dice values of 0.912 (ED) and 0.873 (ES); *ii)* average Hausdorff distances of 7.0 mm (ED) and 7.7 mm (ES) and *iii)* average mean absolute distances of 1.9 mm (ED) and 2.1 mm (ES) [19]. Interestingly, these results are quite close to those obtained by Pedrosa *et al.* [5]. Additionally, the use of only 15 patients during the training phase illustrates the strong potential of deep learning techniques to analyze echocardiographic images.

III. CAMUS Dataset

A. Dataset

As described in the previous section, only one echocardiographic dataset, composed of 45 3D sequences, has been broadly validated by the community. This dataset is not appropriate to study the behavior of deep learning approaches in the particular case of 2D sequences. In this context, we introduce the largest publicly-available and fully annotated dataset for the purpose of 2D echocardiographic assessment.

1) Patient Selection: The proposed dataset consists of clinical exams from 500 patients, acquired at the University Hospital of St Etienne (France) and included in this study within the regulation set by the local ethical committee of the hospital. The acquisitions were optimized to perform LV_{EF} measurements. In order to enforce clinical realism, neither prerequisite nor data selection have been performed. Consequently, *i)* some cases were difficult to trace; *ii)* the dataset involves a wide variability of acquisition settings; *iii)* for some patients, parts of the wall were not visible in the images; *iv)* for some cases, the probe orientation recommendation to acquire a rigorous four-chambers view was simply impossible to follow and a five-chambers view was acquired instead. This produced a highly heterogeneous dataset, both in terms of image quality and pathological cases, which is typical of daily clinical practice data. Table I provides the main information which characterizes the collected dataset. From this table, one can see that half of the dataset population has a LV_{EF} lower than 45%, thus being considered at pathological risk (beyond the uncertainty of the measurement). Also, 19% of the images have a poor quality (based on the opinion of one expert O_{1a}), indicating that for this subgroup the localization of the LV_{Endo} and LV_{Epi} as well as the estimation of clinical indices are not considered clinically accurate and workable. In classical analysis, poor quality images are usually removed from the dataset because of their clinical uselessness. Therefore, those data were not involved in this project during the computation of the different metrics but were used to study their influence as part of the training and validation sets for deep learning techniques.

TABLE I
THE MAIN CHARACTERISTICS OF THE CAMUS ECHOCARDIOGRAPHIC DATASET COLLECTED FROM 500 PATIENTS

Dataset	**Image Quality** *(in percentage)*			**LV_{EF}** *(in percentage)*		
	Good	Medium	Poor	$\leq 45\%$	$\geq 55\%$	else
Full	35	46	19	49	19	32
fold 1	34	48	18	48	20	32
fold 2	34	46	20	50	18	32
fold 3	34	46	20	48	20	32
fold 4	34	46	20	50	20	30
fold 5	34	46	20	48	20	32
fold 6	36	46	18	50	20	30
fold 7	36	46	18	50	20	30
fold 8	36	46	18	50	18	32
fold 9	36	46	18	48	20	32
fold 10	36	46	18	50	18	32

The dataset was divided into 10 folds to perform standard cross-validation for the machine learning methods. Each fold contains 50 patients with the same distributions in terms of image quality and LV_{EF} as the full dataset (see table I). For each of the 10 test sets, the remaining 450 patients (9 folds) were used during the training/validation phases of the machine learning techniques. In particular, 8 folds (400 patients) were used for training and 1 (50 patients) for validation, *i.e.* parameters optimization. The full dataset is available for download at https://camus.creatis.insa-lyon.fr/challenge/.

2) Acquisition Protocol: The full dataset was acquired from GE Vivid E95 ultrasound scanners (GE Vingmed Ultrasound, Horten Norway), with a GE M5S probe (GE Healthcare, US). No additional protocol than the one used in clinical routine was put in place. For each patient, 2D apical four-chamber and two-chamber view sequences were exported from EchoPAC analysis software (GE Vingmed Ultrasound, Horten, Norway). These standard cardiac views were chosen for this study to enable the estimation of LV_{EF} values based on the Simpson's biplane method of discs [21]. Each exported sequence corresponds to a set of B-mode images expressed in polar coordinates. The same interpolation procedure was used to express all sequences in Cartesian coordinates with a unique grid resolution, *i.e.* $\lambda/2 = 0.3$ mm along the x-axis (axis parallel to the probe) and $\lambda/4 = 0.15$ mm along the z-axis (axis perpendicular to the probe), where λ corresponds to the wavelength of the ultrasound probe. At least one full cardiac cycle was acquired for each patient in each view, allowing manual annotation of cardiac structures at ED and ES.

B. Reference Segmentation and Contouring Protocol

Establishing a well-defined ground-truth segmentation was of utmost importance for this work. The main difficulty when delineating 2D echocardiographic images comes from poor contrast in some regions along with the presence of well-known artifacts (*e.g.* reverberation, clutter, acoustic shadowing). One direct consequence is that embedded fully-automatic ultrasound cardiac segmentation softwares do not perform well. During the clinical exam, the clinicians delineate the different contours using semi-automatic tools under time constraints. In this context, the use of manual annotations extracted from clinical exams is not optimal to design

a reference dataset for machine learning where the coherence and accuracy in the manual contouring play an important role during the learning phase. To illustrate this point, it is interesting to note that the only existing shared echocardiographic dataset with reference annotations was realized off-line for the purpose of the CETUS MICCAI challenge in 2014. In particular, it required more than 10 months for 3 different cardiologists to manually contour the 3D endocardium surfaces of 45 patients including consensus revisions [3]. This illustrates the extreme difficulty in designing such a high-quality dataset.

1) Cardiologists Involvement: Three cardiologists (O_1, O_2 and O_3) participated in the annotation of the dataset. Considerable effort was spent to define a consistent manual segmentation protocol. This protocol was designed with the help of O_1 and was then strictly followed O_2 and O_3. In particular, we asked O_1 to perform the manual annotation and to determine ED and ES of the full dataset, while the two others contoured the test set of fold 5 (50 patients). O_1 also annotated twice fold 5 seven months apart (we call those annotations O_{1a} and O_{1b}). This fold was therefore used to measure both the inter- and intra-observer variability.

2) Contouring Protocol: According to the recommendation of the American Society of Echocardiography and the European Association of Cardiovascular Imaging [22], ED is preferably defined as the first frame after mitral valve closure or the frame in the cardiac cycle in which the respective LV dimension or volume measurement is the largest. ES is best defined as the frame after aortic valve closure (*e.g.* using an apical long axis view image) or the frame in which the cardiac dimension or volume is the smallest. In this work, ED and ES were selected as the frames where the LV dimension was at its largest or smallest, which is not the most accurate way, especially in the presence of abnormalities. This simpler approach was used due to the lack of reliable ECG. Thus the clinical indices, ED/ES volume and EF, reported in this work have to be interpreted with this in mind. While only the extraction of the LV_{Endo} contour is necessary to estimate LV_{EF} values, we also asked the cardiologists to manually outline the LV_{Epi} and the LA for all patients. This was done to study the influence of contextualization (segmentation of several structures at once) on the performance of the LV_{Endo} segmentation using deep learning techniques. The following protocol was set up.

- LV_{Endo}: Convention was used for the LV wall, mitral valve plane, trabeculations, papillary muscles and apex [22]. Basic points were to *i)* include trabeculae and papillary muscles in the LV cavity; *ii)* keep tissue consistency between ED and ES instants; *iii)* terminate the contours in the mitral valve plane on the ventricular side of the bright ridge, at the points where the valve leaflets are hinging; *iv)* partially exclude left ventricular outflow tract from the cavity by drawing from septal mitral valve hinge point to the septal wall to create a smooth shape.
- LV_{Epi}: There is no recommendation for delineating the epicardium. We thus outlined the epicardium as the interface between the pericardium and the myocardium for the anterior, anterolateral and inferior segments and the frontier between the right ventricle cavity and the septum for the inferoseptal segments.
- LA: There are recommendations for LA segmentation to assess the full LA area from dedicated LA recordings. However, since we have used acquisitions focusing on the LV, part of the dataset does not cover the full LA surface and is thus not suited to perform such measurement. Having this in mind, we used the following contouring protocol: *i)* start the LA contour from the extremities of the LV_{Endo} contour, at the points where the valve leaflets are hinging; *ii)* have the contour pass by the LA inner border.

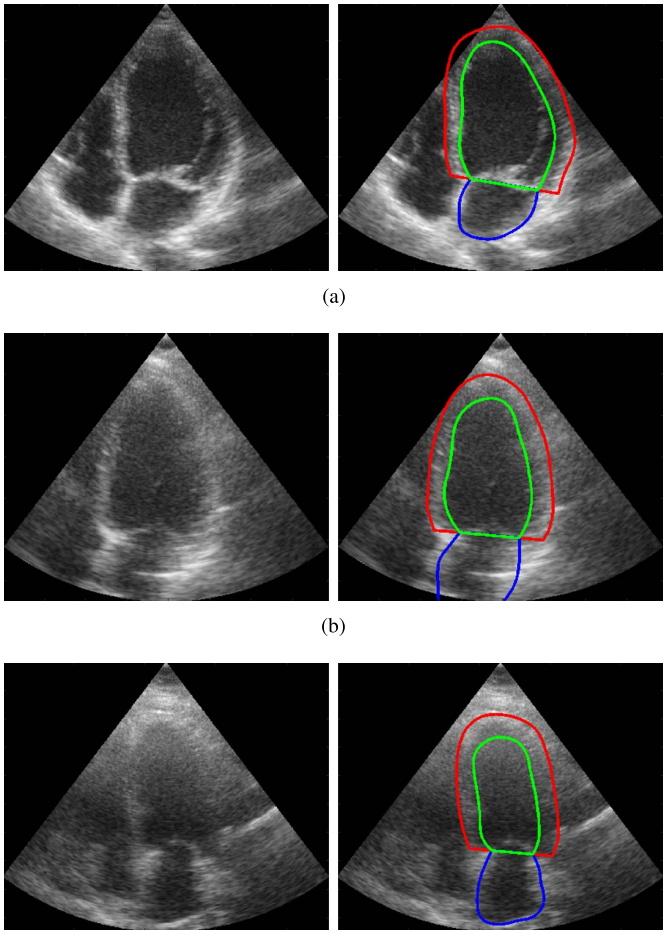

Fig. 1. Typical images extracted from the proposed dataset. Endocardium and epicardium of the left ventricle and left atrium wall are shown respectively in green, red and blue. [Left] input images; [Right] corresponding manual annotations. (a) Good image quality. (b) Medium image quality. (c) Poor image quality.

Fig. 1 illustrates our manual contouring protocol for a good, a medium, and a poor-quality image.

IV. Evaluated Methods

A. CNN Techniques Based on an Encoder-Decoder Architecture

The goal of this study is to assess how far CNNs can go at segmenting 2D echocardiographic images. As such, we chose to focus on the well-known encoder-decoder networks (EDNs) which have been the cornerstone of a wide variety of CNNs

TABLE II
MAIN CHARACTERISTICS OF U-NET 1 AND U-NET 2. MORE DETAILS ARE PROVIDED IN THE SUPPLEMENTARY MATERIALS/MULTIMEDIA TAB

Architectures	Number of feature maps	Lowest Resolution	Upsampling Scheme	Normalization Scheme	Batch Size	Learning Rate	Loss Function	# Trainable Parameters
U-Net 1	$32 \downarrow 128 \uparrow 16$	$8 * 8$	$2 * 2$ repeats	None	32	1e-3	Multi-class Dice	2M
U-Net 2	$32 \downarrow 512 \uparrow 32$	$16 * 16$	Deconvolutions	BatchNorm	10	1e-4	Cross-entropy + weight decay	18M

that were successfully applied in medical imaging [23]. EDNs are based on a two-stage convolutional network architecture well suited for segmentation. The first part, known as the encoder, consists of a series of convolutions and downsampling operations. These operations extract features from the images while spatially compressing them, thus enabling extraction of high-level features. The second part is the decoder, which uses features from the encoder and applies a set of convolutions and upsampling operations to gradually transform feature maps into a final segmentation.

Among the existing EDNs, one of the most popular architectures used in medical imaging corresponds the U-Net model proposed by Ronneberger *et al.* [18] in 2015. This network integrates skip connections between the encoder and decoder parts with the goal of helping in retrieving details that were potentially lost during the downsampling while also stabilizing gradients. The original U-Net follows a specific scheme of convolutions, where each downsampling and upsampling step is proceeded by two 3x3 convolutional layers, while the number of features is doubled per downsampling and reduced in half per upsampling. U-Net has been successfully applied to a wide range of medical applications [23], but for each application, the network design has usually been adapted and optimized to get the best segmentation performance on each application. The main U-Net design choices can be classified in three categories: *i)* layer choices: convolutional layer size, activation functions, normalization layers, down- and upsampling strategies (*e.g.* max pooling, striding, deconvolution and repeat); *ii)* the optimization process (gradient descent strategy, weight initialization, loss function, batch size, regularization constraints, stopping criteria, deep supervision, dropout); *iii)* data handling (pre-processing, augmentation, sampling). Since the seminal paper in 2015, several studies based on the EDN structure have been carried out with the goal of outperforming the U-Net. Among those methods, two types of approaches have been proposed: those based on U-Net architecture but with extensions such as shape regularization [19] and those with more sophisticated architectures [24], [25]. In this context, we decided to benchmark the following EDNs for the purpose of segmenting 2D echocardiographic images:

1) U-Net: Taking into account the wide range of possible U-Net designs, we decided to compare the performance of two independent implementations, *i.e.* U-Net 1 optimized for speed, and U-Net 2 optimized for accuracy. This leads to two different architectures (which both differ from the original one proposed by Ronneberger *et al.*) with their own hyperparameters settings, as shown in table II. The "number of feature maps" column given in table II corresponds to the number of convolutions in the convolution layers. For each U-Net implementation, we successively indicate the values for the first, the bottom (where the spatial information is the most compressed), and the last convolution layers. U-Net 1 & 2 enable to investigate the impact of hyperparameters choices on the quality of the results. The total number of parameters for U-Net 1 and 2 are 2.0M and 17.5M, respectively.

2) ACNN: Starting from a given segmentation architecture, this method integrates an auxiliary loss to constrain the segmentation output to fit a non-linear compact representation of the underlying anatomy derived from an auto-encoder network [19]. For comparison purposes, we used the U-Net 1 architecture described in table II as the segmentation module in our ACNN implementation. Moreover, the following choices were made to obtain the best results on our dataset: *i)* a code of 32 coefficients was set for the auto-encoder network (which allows an average reconstruction accuracy of 97%); *ii)* the hyperparameter balancing the segmentation and shape regularization losses was set so that the two losses had close initialization values. The ACNN models in our study have 2.2M parameters.

3) SHG: Stacked Hourglasses (SHG) method integrates three successive encoder-decoder networks (usually three times the same architecture) where the first two are used as residual blocks [24]. Each output of the encoder-decoder networks is associated with an intermediate segmentation loss. This strategy is called deep supervision. The output of the third network is used as the final segmentation result. For comparison purposes, we also used the U-Net 1 architecture as the key encoder-decoder network in our SHG implementation. The number of parameters of the SHG method is 4.5M (not 6M in order to keep the same batch size as U-Net 1).

4) U-Net++: This method is also based on deep supervision technique but with the integration of additional convolution layers in the form of dense skip connections [25]. Starting from the official online version of the code, we adapted the corresponding architecture to obtain the best results on our dataset. The following changes were made: *i)* dropout was removed; *ii)* averaging of the last feature maps of the intermediate outputs was removed; *iii)* the original design of layers was adapted according to the choices we made to optimized U-Net 1 architecture; *iv)* the batch size was set to 20. The U-Net++ method comprises 1.1M parameters. The original version had 9M parameters but we adapted it, in particular the number of feature maps, so to get the best possible results on the CAMUS dataset.

Please note that the same data pre- and post-processing strategies (*i.e.* connected component analysis keeping the largest region and removing holes) were applied for each of the evaluated method. For completeness' sake, implementation details of the two U-Net architectures as well as additional tests on ACNN and U-Net++ can be found in the supplementary materials.

B. Non-Deep Learning State-of-the-Art Techniques

To compare the performance of the EDN methods described above, we implemented the following non-deep learning state-of-the-art methods which obtained among the best results during the CETUS challenge [3] and which were recently improved [5] and applied in 2D [26].

1) SRF: Structured Random Forests (SRF) refer to an ensemble learning method for classification or regression. It operates at training time by building a set of decision trees that assign a label patch to each input image patch, computed as the mean prediction of the individual trees [27]. During the training phase, each tree individually learns a set of split functions from a random subset of the training dataset and input features. Those functions are intended to group patches sharing close image intensities and segmentation patterns. During the testing phase, the image to segment is fragmented into different overlapping patches. Each image patch goes through the splitting functions of each tree so that the mean label patch computed from the reached leaves forms its segmentation. Detailed description of the SRF algorithm implemented in this project can be found in [26]. Compared to our previous study, data was not split between ES and ED nor between 4 chambers and 2 chambers views but processed in one indistinctive pool of images. Since CAMUS has a larger number of patients than the dataset used in [26], we trained 12 individual trees for each subset of 100 patients.

2) BEASM: The key concept of the B-Spline Explicit Active Surface Model (BEASM) framework is to consider the boundary of a deformable interface as an explicit function, where one of the coordinates of the points within the surface is given explicitly as a function ψ of the remaining coordinates. In this framework, ψ is defined as a linear combination of B-spline basis functions whose controlled knots are located on a regular rectangular grid defined on the chosen coordinate system (polar space in our case). Based on a standard variational approach, the evolution of the deformable surface is then governed by the minimization of an energy function according to the B-spline coefficients [28]. This framework has been successfully applied in [29] for the coupled segmentation of the $\mathrm{LV_{Endo}}$ and $\mathrm{LV_{Epi}}$ structures in echocardiography and further extended by the integration of a shape prior directly into the B-spline space in [5], named as BEASM in the rest of the paper. Because BEASM amounts to a deformable-based model, the initialization of the contour plays a crucial role on the quality of the results. We thus decided to implement two different strategies: *i)* one named BEASM-fully where the evolving contour is automatically initialized from a method inspired by the work proposed in [30]; *ii)* another named BEASM-semi where the evolving contour is initialized from three points (two at the base and one at the apex of the $\mathrm{LV_{Endo}}$ structure) extracted from the reference contours. By doing so, we gave the possibility to quantify the influence of the initialization procedure for BEASM on an heterogeneous ultrasound dataset.

V. Results

As stated in Section III-A, the 19% of poor quality images were not used to compute the different metrics provided in this part. Moreover, to avoid the use of different models according to the acquisition settings, we trained only one model for each machine learning method on both apical four-chamber and two-chamber views regardless of the time instant in the cardiac sequence. For a detailed analysis of the results, a set of complementary information are provided in the supplementary materials (*i.e.* figures of segmentation results for all the methods, Bland-Altman plots, limitations examples).

A. Evaluation Metrics

1) Geometrical Metrics: To measure the accuracy of the segmentation output ($\mathrm{LV_{Endo}}$, $\mathrm{LV_{Epi}}$ or LA) of a given method, the Dice metric, the mean absolute distance (d_m) and the 2D Hausdorff distance (d_H) were used. The Dice similarity index is defined as $D = 2\left(\left|S_{user} \cap S_{ref}\right|\right) / \left(\left|S_{user}\right| + \left|S_{ref}\right|\right)$ and is a measure of overlap between the segmented surface S_{user} extracted from a method and the corresponding reference surface S_{ref}. The Dice index gives a value between 0 (no overlap) and 1 (full overlap). d_m corresponds to the average distance between S_{user} and S_{ref} while d_H measures the local maximum distance between the two surfaces.

2) Clinical Metrics: We gauge the methods' performance with 3 clinical indices: *i)* the ED volume ($\mathrm{LV_{EDV}}$ in ml); *ii)* the ES volume ($\mathrm{LV_{ESV}}$ in ml); *iii)* the ejection fraction ($\mathrm{LV_{EF}}$ as a percentage) for which we computed four metrics: the correlation ($corr$), the bias and the standard deviation (std) values (computed from conventional definitions) and the mean absolute error (mae). The combination of the bias and standard deviation also provides useful information on the corresponding limit of agreement values.

B. Empirical Results

1) Geometrical Scores: Table III shows the segmentation testing accuracy computed from patients having good and medium image quality (406 patients) for the 8 algorithms described in section IV. Mean and standard deviation values for each metric were obtained from cross-validating on the 10 folds of the dataset. The values in bold correspond to the best scores for each metric. From these results, one can see that the EDN implementations get the overall best segmentation scores on all metrics, for both ED and ES. Interestingly, while the EDN methods are fully-automatic, they still get better segmentation results than the semi-automatic BEASM algorithm.

TABLE III

SEGMENTATION ACCURACY (LV_{Endo} AND LV_{Epi}) OF THE 8 EVALUATED METHODS ON THE TEN TEST FOLDS OF TABLE I RESTRICTED TO PATIENTS HAVING GOOD & MEDIUM IMAGE QUALITY (406 PATIENTS IN TOTAL). ALL THE METRICS WERE COMPUTED USING THE ANNOTATIONS OF EXPERT O_{1a}. ALL THE SCORES OBTAINED WITH THE U-NET 2, WHICH IS THE OVERALL BEST PERFORMING METHOD, WERE STATISTICALLY DIFFERENT WITH A p-VALUE < 0.05 (COMPUTED WITH A WILCOXON SIGNED-RANK TEST) COMPARED TO ALL THE TESTED METHODS, EXCEPT U-NET 1 AND SHG FOR THE d_H METRIC

Methods *	ED						ES					
	LV_{Endo}			LV_{Epi}			LV_{Endo}			LV_{Epi}		
	D	d_m	d_H	D	d_m	d_H	D	d_m	d_H	D	d_m	d_H
	val.	mm	mm	val.	mm	mm	val.	mm	mm	val.	mm	mm
O_{1a} vs O_2 (inter-obs)	0.919 ±0.033	2.2 ±0.9	6.0 ±2.0	0.913 ±0.037	3.5 ±1.7	8.0 ±2.9	0.873 ±0.060	2.7 ±1.2	6.6 ±2.4	0.890 ±0.047	3.9 ±1.8	8.6 ±3.3
O_{1a} vs O_3 (inter-obs)	0.886 ±0.050	3.3 ±1.5	8.2 ±2.5	0.943 ±0.018	2.3 ±0.8	6.5 ±2.6	0.823 ±0.091	4.0 ±2.0	8.8 ±3.5	0.931 ±0.025	2.4 ±1.0	6.4 ±2.4
O_2 vs O_3 (inter-obs)	0.921 ±0.037	2.3 ±1.2	6.3 ±2.5	0.922 ±0.036	3.0 ±1.5	7.4 ±3.0	0.888 ±0.058	2.6 ±1.3	6.9 ±2.9	0.885 ±0.054	3.9 ±1.9	8.4 ±2.8
O_{1a} vs O_{1b} (intra-obs)	0.945 ±0.019	1.4 ±0.5	4.6 ±1.8	0.957 ±0.019	1.7 ±0.9	5.0 ±2.3	0.930 ±0.031	1.3 ±0.5	4.5 ±1.8	0.951 ±0.021	1.7 ±0.8	5.0 ±2.1
SRF	0.895 ±0.074	2.8 ±3.6	11.2 ±10.2	0.914 ±0.057	3.2 ±2.0	13.0 ±9.1	0.848 ±0.137	3.6 ±7.8	11.6 ±13.6	0.901 ±0.078	3.5 ±4.7	13.0 ±11.1
BEASM-fully	0.879 ±0.065	3.3 ±1.8	9.2 ±4.9	0.895 ±0.051	3.9 ±2.1	10.6 ±5.1	0.826 ±0.092	3.8 ±2.1	9.9 ±5.1	0.880 ±0.054	4.2 ±2.0	11.2 ±5.1
BEASM-semi	0.920 ±0.039	2.2 ±1.2	6.0 ±2.4	0.917 ±0.038	3.2 ±1.6	8.2 ±3.0	0.861 ±0.070	3.1 ±1.6	7.7 ±3.2	0.900 ±0.042	3.5 ±1.7	9.2 ±3.4
U-Net 1	0.934 ±0.042	1.7 ±1.0	5.5 ±2.9	0.951 ±0.024	1.9 ±0.9	5.9 ±3.4	0.905 ±0.063	1.8 ±1.3	5.7 ±3.7	0.943 ±0.035	2.0 ±1.2	6.1 ±4.1
U-Net 2	**0.939** ±0.043	**1.6** ±1.3	**5.3** ±3.6	**0.954** ±0.023	**1.7** ±0.9	6.0 ±3.4	**0.916** ±0.061	**1.6** ±1.6	**5.5** ±3.8	**0.945** ±0.039	**1.9** ±1.2	6.1 ±4.6
ACNN	0.932 ±0.034	1.7 ±0.9	5.8 ±3.1	0.950 ±0.026	1.9 ±1.1	6.4 ±4.1	0.903 ±0.059	1.9 ±1.1	6.0 ±3.9	0.942 ±0.034	2.0 ±1.2	6.3 ±4.2
SHG	0.934 ±0.034	1.7 ±0.9	5.6 ±2.8	0.951 ±0.023	1.9 ±1.0	**5.7** ±3.3	0.906 ±0.057	1.8 ±1.1	5.8 ±3.8	0.944 ±0.034	2.0 ±1.2	**6.0** ±4.3
U-Net ++	0.927 ±0.046	1.8 ±1.1	6.5 ±3.9	0.945 ±0.026	2.1 ±1.0	7.2 ±4.5	0.904 ±0.060	1.8 ±1.0	6.3 ±4.2	0.939 ±0.034	2.1 ±1.1	7.1 ±5.1

* *LV_{Endo}: Endocardial contour of the left ventricle; LV_{Epi}: Epicardial contour of the left ventricle; ED: End diastole ES: End systole; D: Dice index; d_m: mean absolute distance; d_H: Hausdorff distance; mae: mean absolute error*
The values in bold refer to the best performance for each measure.
The inter and intra-observer measurements were computed from fold 5 restricted to patients having good & medium image quality (40 patients)

The two U-Nets achieve equivalent results for all the metrics compared to the ones obtained by the more sophisticated encoder-decoder architectures. This hints to the idea that a plateau has been reached, which classical tuning, shape regularization techniques and more sophisticated architectures have difficulties to overcome. This also suggests that a U-Net implementation, which requires less parameters than SHG and U-Net++ methods and less training time than ACNN, offers the best compromise between the network size and performance for the particular task of 2D echocardiographic image segmentation.

To assess the influence of the layer design in the performance between U-Net 1 and 2, statistical significance of their respective results was analyzed by performing the Wilcoxon signed-rank test for each metric. Results showed that U-Net 1 and 2 produced scores that are statistically different (p-value < 0.05) for most metrics at ED and ES, apart for the LV_{Epi} Hausdorff distance. However, this must be nuanced by the fact that *i)* the U-Net geometrical scores are very close (mean d_m and d_H difference of 0.1 mm and 0.1 mm, respectively), producing distributions with high degree of overlap as shown in the supplementary materials; *ii)* the U-Net geometrical results lie between the inter-observer and intra-observer scores for all metrics, proving the robustness of this method in obtaining accurate segmentation results. As a complement to the above, we investigated in the supplementary materials/multimedia tab the influence of each of the U-Net hyperparameters presented in table II. Results highlight the importance of the choice of the normalization scheme.

As for the fully automatic non-deep learning state-of-the-art methods, BEASM-auto obtained on average better Hausdorff distances (mean d_H of 9.9 mm at ED and 10.5 mm at ES) while the SRF got better Dice and d_m scores (mean d_m of 3.0 mm at ED and 3.5 mm at ES). However, the large standard deviation values for the SRF illustrate the difficulties of this method in obtaining consistent segmentations over the entire

TABLE IV

CLINICAL METRICS OF THE 8 EVALUATED METHODS ON THE TEN TEST FOLDS OF TABLE I RESTRICTED TO PATIENTS HAVING GOOD & MEDIUM IMAGE QUALITY (406 PATIENTS IN TOTAL). ALL THE METRICS WERE COMPUTED USING THE ANNOTATIONS OF EXPERT O_{1a}. VOLUMES AND EJECTION FRACTION OBTAINED WITH U-NET 2 WERE STATISTICALLY DIFFERENT WITH p-VALUES < 0.05 (COMPUTED WITH THE WILCOXON SIGNED-RANK TEST) COMPARED TO ALL THE TESTED METHODS, EXCEPT SHG FOR THE LV_{ESV} INDICE

Methods *	LV_{EDV}			LV_{ESV}			LV_{EF}		
	corr	***bias*** $\pm\sigma$	***mae***	***corr***	***bias*** $\pm\sigma$	***mae***	***corr***	***bias*** $\pm\sigma$	***mae***
	val.	ml	ml	val.	ml	ml	val.	%	%
O_{1a} vs O_2 (inter-obs)	0.940	18.7±12.9	18.7	0.956	18.9±9.3	18.9	0.801	-9.1±8.1	10.0
O_{1a} vs O_3 (inter-obs)	0.895	39.0±18.8	39.0	0.860	35.9±17.1	35.9	0.646	-12.6±10.0	13.4
O_2 vs O_3 (inter-obs)	0.926	-20.3±15.6	21.0	0.916	-17.0±13.5	17.7	0.569	3.5±11.0	8.5
O_{1a} vs O_{1b} (intra-obs)	0.978	-2.8±7.1	6.2	0.981	-0.1±5.8	4.5	0.896	-2.3±5.7	0.9
SRF	0.755	-0.2±25.7	17.4	0.827	9.3±18.0	14.8	0.465	-11.5±15.4	12.8
BEASM-fully	0.704	13.4±30.6	22.9	0.713	18.0±25.8	22.5	0.731	-9.8±8.3	10.7
BEASM-semi	0.886	14.6±19.2	17.8	0.880	18.3±16.9	19.5	0.790	-9.4±7.2	10.0
U-Net 1	0.947	-8.3±12.6	10.9	0.955	-4.9±9.9	8.2	0.791	-0.5±7.7	5.6
U-Net 2	**0.954**	-6.9 ±11.8	**9.8**	**0.964**	-3.7± 9.0	**6.8**	**0.823**	-1.0±7.1	**5.3**
ACNN	0.945	-6.7± 12.9	10.8	0.947	-4.0± 10.8	8.3	0.799	-0.8±7.5	5.7
SHG	0.943	6.4±12.8	10.5	0.938	-3.2±11.3	8.2	0.770	-1.4±7.8	5.7
U-Net ++	0.946	-11.4± 12.9	13.2	0.952	-5.7± 10.7	8.6	0.789	-1.8± 7.7	5.6

* *corr: Pearson correlation coefficient; mae: mean absolute error.*
The values in bold refer to the best performance for each measure.
The inter and intra-observer measurements were computed from fold 5 restricted to patients having good & medium image quality (40 patients.)

dataset. As for the BEASM-semi, one can see that the manual initialization has a strong impact on the quality of the results, with a mean improvement of 0.8 mm and 2.4 mm for the d_m and d_H metrics, respectively. Moreover, it is well known that the left ventricle shape is more difficult to segment at ES, leading to slightly worse performance for classical algorithms on this time instant. This property is also confirmed in our study since all the evaluated methods produced better results at ED on every metric.

As complement, we provided in the supplementary materials/multimedia tab the geometrical scores obtained on the poor quality images (94 patients) for the 8 evaluated algorithms. For this part of the dataset, the EDNs also obtained the best segmentation results on all metrics. Interestingly, while EDN scores on poor quality images are slightly worse than those computed on good and medium quality, they remain very competitive compared to the scores given in table III (mean LV_{Endo} d_m and d_H of 2.2 mm and 7.0 mm and mean LV_{Epi} d_m and d_H of 2.3 mm and 7.6 mm).

2) Clinical Scores: Table IV contains the clinical metrics for the 8 methods. Those indices were computed with the Simpson's rule [21] from the segmentation results of each algorithm. The values in bold represent the best scores for the corresponding index. As for segmentation, the EDNs obtained the best clinical scores on all the tested metrics (bias was not taken into account since the lowest bias value in itself does not necessarily mean the best performing method). Regarding the estimation of the LV_{EDV} and LV_{ESV}, the EDNs obtained high correlation scores (all above 0.94) and reasonably small biases (at most 11.4 ml), standard deviations (less than 12.9 ml) and mean absolute errors (at most 13.2 ml). Results are more contrasted for the estimation of the LV_{EF}. For this metric, the EDNs got lower correlation scores (at most 0.82) but smaller biases (less than 1.8 %), standard deviations (at most 7.8 %) and mean absolute errors (less than 5.7 %). It is worth pointing out that average EDN scores are all below the inter-observer scores. This proves the clinical interest of such approaches but also reveals the needs for improvement as discussed in Section V-D. Here again, even if the U-Net methods involved simpler architecture, they obtained similar results compared to the more sophisticated EDNs. Finally, using the Wilcoxon signed-rank test, U-Net 1 and U-Net 2 produced LV_{EDV}, LV_{ESV} and LV_{EF} results whose difference is statistically significant (p-values < 0.05), although their measurements are very close.

C. U-Net Behavior

From the results given in table III and IV, it appears that the U-Net method has the most effective architecture among the tested EDN models in terms of trade-off between the number of parameters and the achieved performance for the particular task of 2D echocardiographic image analysis. To better analyze the behavior of this model, we set up several additional experiments whose results are provided in Fig. 2 and Fig. 3. For all these experiments, even if the acquisitions were optimized to perform LV_{EF} measurements (meaning that part of the LA may or may not be fully visible depending on the acquisitions), we also investigated the capacity of U-Net to segment the LA in addition to the LV_{endo} and LV_{epi}. Moreover, since the two tested U-Nets produced overall close geometrical and clinical scores, we only used in this part the U-Net 1 model since it requires considerably less parameters to learn. Finally, all the given metrics were computed from both four and two-chamber views and at ED and ES time instants to facilitate the interpretation of the results.

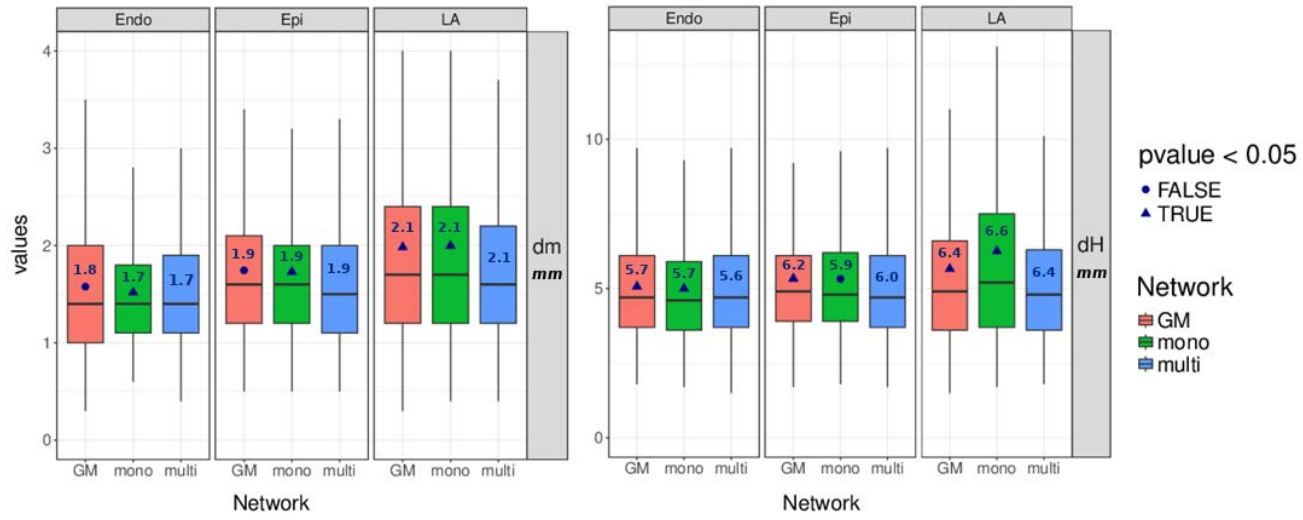

Fig. 2. Tukey box plots computed from the geometrical results of the U-Net 1 architecture for three different schemes (*GM* for learning to simultaneously segment all three structures from good & medium image quality, *mono* for learning to segment one structure from all image quality, *multi* for learning to simultaneously segment all three structures from all image quality). Blue numbers correspond to mean values computed from each set of measurements. All p-values are based on the Wilcoxon signed-rank test computed with the *multi* strategy as reference.

1) Mono Versus Multi-Structures Approaches: We assessed the influence of learning strategies on the quality of the segmentation of the LV_{Endo}, LV_{Epi} and LA. In particular, we trained 4 models with the same U-Net 1 architecture but with different training sets including all image quality, *i.e.* one network trained on predicting only the LV_{Endo}, one the LV_{Epi}, one the LA, and one all structures. Results on the full dataset are plotted in green and blue in Fig. 2 and are referred to as *mono* and *multi*.

From the derived box plots, one can see that, unrelated to the structure, the mono and multi-structures approaches produced very close results even if the corresponding differences are statistically different. These results show that, with the proposed implementations, learning the segmentation of one structure (*e.g.* LV_{Endo}) in the context of the others (*e.g.* LV_{Epi} & LA) does not improve significantly the results compared to learning the segmentation of the structure alone. This hints at designing dedicated architectures and/or loss functions to better exploit the contextual information provided in the segmentation masks. Furthermore, even if the segmentation of the LA structure is challenging compared to LV_{Epi} and LV_{Endo} due to acquisition conditions, the U-Net 1 manages to get close results both in terms of mean absolute distance (mean d_m equals to 1.7, 1.9 and 2.1 mm for the LV_{Endo}, LV_{Epi} and LA respectively) and average Hausdorff distance (mean d_H equals to 5.6, 6.0 and 6.4 mm for the LV_{Endo}, LV_{Epi} and LA respectively).

2) The Effect of Poor Quality Images: We investigated in Fig. 2 the influence of involving images of poor quality during the training phase. Based on a multi-structures scheme, we trained two U-Net 1 models with the same architecture, one using the full training dataset not caring for image quality (plotted in blue and referred as *multi* in Fig. 2) and one using the training dataset restricted to patients having good and medium image quality (plotted in red and referred as *GM* in Fig. 2). From the obtained box plots, one can observe that the two different strategies produced very close results even if the corresponding differences are mostly statistically significant (apart for the d_m metric for the LV_{Endo} and LV_{Epi}). These results suggest that the 19% (94 patients) of poor quality images *i)* do not bring additional information (supporting that the remaining deep learning issues are weakly linked to image quality); *ii)* do not decrease performance compared to a model trained on the 406 patients with good and medium image quality. This result suggests that poor image quality, in itself, does not complicate the segmentation task as much as could be expected and that encoder-decoder based techniques are able to cope with the variability in image quality found in echocardiography.

3) Influence of the Size of the Training Dataset: We studied in Fig. 3 the influence of the size of the training dataset on the quality of the segmentation of the LV_{Endo}, LV_{Epi} and LA structures. To this aim, we set up 8 different experiments, where the same fold 5 and 6 were respectively used as test and validation sets. As for the training set, starting from 50 patients, we added for each new experiment 50 additional patients until 400 patients was reached for the last trial. In each experiment, the same U-Net 1 architecture was used and optimized in the same way to derive the best performing parameters from the validation set. Moreover, the number of training epochs was proportionally lowered to ensure that each network went through the same number of iterations.

From this figure, one can first observe an overall improvement of all metrics for the three cardiac structures with the increasing number of patients in the training set. Interestingly, while the improvement between 50 to 200 patients is quite pronounced (*e.g.* a decrease of d_H for the LV_{Endo} from 7.2 mm to 5.8 mm), one can observe a change in the evolution of the performance of the U-Net 1 method from 250 patients. Indeed, for this particular value, results worsen a bit, which may be explained by the bias brought by the validation and test data as we are not doing cross-validation in this experiment.

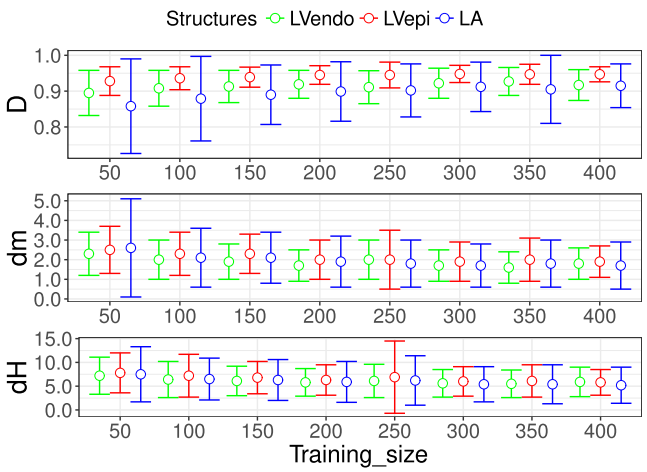

Fig. 3. Evolution of the segmentation scores (computed from fold 5 and from both ED and ES time instants) derived from the U-Net 1 architecture according to an incremental increase of the number of patients involved in the training dataset. Dots correspond to mean values while bars correspond to standard deviations.

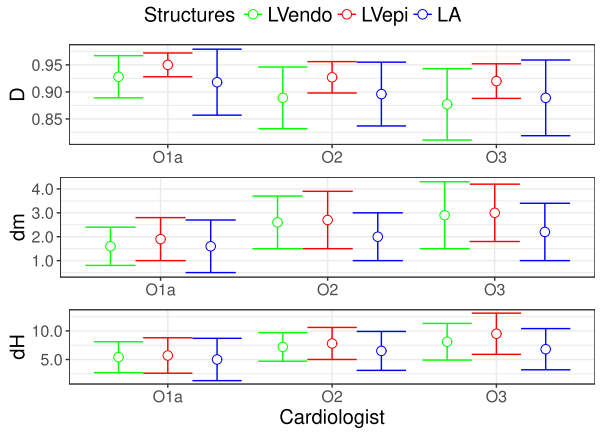

Fig. 4. Geometric scores of the three cardiologist-specific models on 400 patients (1600 images).

Moreover, from this value, the d_m scores seem to stabilize around 1.8 mm for the $\mathrm{LV}_{\mathrm{Endo}}$, 1.9 mm for the $\mathrm{LV}_{\mathrm{Epi}}$ and 1.8 mm for the LA structure. The same conclusions can be made for the Dice metric, with a convergence value around 0.920 for the $\mathrm{LV}_{\mathrm{Endo}}$, 0.947 for the $\mathrm{LV}_{\mathrm{Epi}}$ and 0.909 for the LA structure. As for the d_H metric, while some improvement can still be observed from 250 to 400 patients for the $\mathrm{LV}_{\mathrm{Epi}}$ and LA structures (1.1 mm and 1.0 mm for the $\mathrm{LV}_{\mathrm{Epi}}$ and the LA, respectively), it is not obvious to draw the same conclusion for the $\mathrm{LV}_{\mathrm{Endo}}$ structure since the decrease of its corresponding value is less pronounced (0.2 mm). In the light of these results, the U-Net 1 implementation performs better that the state-of-the-art non-deep learning methods after training with only 50 patients. Moreover, this method needs at least 250 patients during the training phase to reach highly competitive results, which can be slightly improved with a larger training set.

4) Influence of the Expert Annotations: We investigated in Fig. 4 the influence of the expert annotations during the training phase. To this aim, we trained three models on fold 5 from the same U-Net 1 architecture based each time on the manual contouring from a different annotator. The validation fold was kept the same for each experiment to avoid any bias error. The models were then evaluated on the remaining 400 patients annotated by cardiologist O_{1a}.

From this figure, one can observe that the best scores for the three structures are obtained for the model trained on the annotations of cardiologist O_{1a}, who performed the manual contouring on the test and validation sets. This observation is consistent with the inter-variability results provided in tables III and IV. It confirms that cardiologists have consistent differences in their way of contouring images and that an EDN has the capacity to learn a specific way of segmenting.

5) Runtime Performance: The two U-Nets were implemented in Python with the same version of the TensorFlow and Keras libraries and an Nvidia Tesla M60 GPUs (8 Go RAM). Because of the larger number of trainable parameters involved in the U-Net 2 solution (see table II), the running time of the two networks is different. For the training phase, the time required to train on 400 patients is 24 ± 5 *min* and 73 ± 1 *min* for the U-Net 1 and U-Net 2, respectively. At test time, the segmentation of a single image takes 0.09 ± 0.03 *s* and 0.14 ± 0.06 *s* for the U-Net 1 and U-Net 2, respectively.

D. Discussion

1) Statistical Differences Between U-Net 1 and 2 Results: From Tables III and IV it has been observed that although U-Net 1 and U-Net 2 have very similar performances, their results were judged most of the time as being statistically different by the Wilcoxon Signed Rank Test (p-value < 0.05). This can be explained by the fact that when the number of samples is quite high, any slight but consistent deviation between the two distributions will make the difference statistically significant. In our study, since we worked on a large scale dataset, most of the statistical tests were performed on a large amount of samples (for instance Tables III and IV involve more than 800 paired observations for each statistical test), encouraging situations where the differences between results produced by two methods are recognized as statistically significant (even if the evaluated distributions are very close).

2) Inter and Intra-Observer Variability: To further assess the quality of the EDN segmentation results, we added in table III the inter and intra-observer variability measurements computed from fold 5 (restricted to 40 patients with good and medium image quality). Concerning the inter-observer variability, the corresponding Dice scores vary between 0.82 and 0.93, the d_m between 2.2 mm and 4.0 mm and the d_H between 6.0 mm and 8.8 mm. The $\mathrm{LV}_{\mathrm{Epi}}$ is the most difficult structure to annotate while both $\mathrm{LV}_{\mathrm{Endo}}$ and $\mathrm{LV}_{\mathrm{Epi}}$ are harder to contour at ES than at ED. One should also note the large d_m value of 4.0 mm between observer 1a and 3 for the $\mathrm{LV}_{\mathrm{Endo}}$ structure at ES. This illustrates *i)* the difficulty in getting coherent manual annotations between experts from daily clinical practice data; *ii)* the difficulty for the experts to use unfamiliar software for the analysis; *iii)* the needs to provide interactively the volumetric results to the experts for instant comparison (this was not done during the manual annotations); *iv)* the difficulty

in contouring some data acquired with non-standard views. Concerning the intra-observer variability, one can observe that the results obtained on all the segmentation scores are better than the inter-observer ones, with a mean difference of 1.5 mm for the d_m metric and 2.6 mm for the Hausdorff distance. This illustrates the high consistency of manual contouring from experienced cardiologists, even on challenging data. Those results also provide important information on the limits to reach in order to consider that a machine learning algorithm faithfully reproduces the expertise of one cardiologist.

In table IV, we also reported the inter and intra-observer variability measurements computed from fold 5 (restricted to 40 patients with good and medium image quality) for the LV_{EDV}, LV_{ESV} and LV_{EF} metrics. From these results, one can observe that the experts reached good agreements for the estimation of the LV_{EDV} and LV_{ESV} with mean correlation scores of 0.92 and 0.91, respectively. However, the LV_{EF} results are worse with a mean correlation value of 0.67. This reveals the extreme difficulty in getting consistent fully manual annotations from ED to ES and between clinicians. It also illustrates the need for semi- or fully-automatic solutions to get higher temporal coherency, as illustrated by the higher LV_{EF} scores obtained by the semi-automatic BEASM method (0.79) and the EDN approaches (0.79 on average). Concerning the intra-observer variability, results are much more consistent with mean correlation scores of 0.98, 0.98 and 0.90 of the LV_{EDV}, LV_{ESV} and LV_{EF} metrics, respectively.

3) U-Net Versus More Sophisticated Encoder-Decoder Architectures: Tables III and IV underlines that U-Net results are very close to those obtained by more sophisticated architectures. This is surprising as one might expect that more complex deep learning designs would improve results, at least marginally. As for ACNN, similar scores may be explained by the simple shapes encountered in 2D echocardiography. Indeed, the reference contours drawn by the experts involve truncated ellipse-like shapes whose information seems to be easily learned by the different EDNs. As a result, the anatomical constraint of the ACNN does not bring any additional value during the segmentation process, leading to similar or even slightly lower performance due to the regularization effect (which can lead to simpler shapes than expected). Further results which support this hypothesis are provided in the supplementary materials. Concerning SHG and U-Net++, the similar scores may be explained by the results in Fig. 3. From this figure, we observed that U-Net reaches a plateau in terms of its performances when training on more than 250 patients. This suggests that the capacity of a U-Net is sufficient to generalize well on CAMUS dataset. Thus more complex architectures like SHG and U-Net++ did not bring any improvement in our case.

4) Effect of the Stochasticity During the Training Phase: The training of all models in this work is stochastic, thus training a network will not converge to the exact same model each time. To estimate its effect on performance, we provide in table III of the supplementary materials/multimedia tab the results produced by U-Net 1 and U-Net 2 for two different trainings, showing that at worst the d_m and d_H respectively varied of 0.1 and 0.2 mm. The scores obtained with the two U-Net 2 models were consistently better than the ones produced by the U-Net 1 models. This indicates that the effect of the stochastic nature of the training process is limited in our case.

5) Accuracy of EDNs at Delineating the LV_{Endo}, LV_{Epi} and LA Structures: Segmentation results given in table III show that the five EDN implementations clearly outperform the state-of-the-art fully and semi-automatic non-deep-learning methods. In particular, while also learning from annotated data, SRF does not perform as consistently as the EDNs. Concerning the deformable model-based BEASM, even if it integrates the annotated information through a shape prior, this method produces overall significantly less accurate segmentation results. It thus appears from this study that a well-designed EDN can reach impressive segmentation scores in echocardiography. In addition, results given in Section V-C show that U-Net provides the same performances whether low quality data is included or not, and whether the model learns to segment all structures simultaneously or separately (involving three models instead of a single one). U-Net 1 obtained very close results compared to U-Net 2, but with 9 times less parameters. The number of parameters directly influences runtime performance, training time, storage and memory consumption. Since ultrasound is a real-time imaging modality, it would be a huge advantage if a single compact EDN could accurately segment multiple cardiac structures. In this regard, this study indicates that U-Net 1 would be the best candidate to embed into clinical equipment.

Interestingly, the EDN results are better than the inter-observer scores, on all structures and metrics. Although further investigations shall be made to validate this assertion, the obtained results tend to show that, when properly trained, deep learning techniques are able to reproduce manual annotations with high fidelity. The results presented in this pilot study should thus stimulate the community to set up public multi-centric and multi-vendor datasets in echocardiography with annotations from cardiologists having passed high level consensus criteria. It is also interesting to note that EDN results are slightly worse than the intra-observer scores, on all structures and metrics (apart for the d_m metric for the LV_{Epi} at ED). This reveals that even if EDNs produce remarkable results, there still exists room for improvement to faithfully reproduce the manual annotations of one expert, taking into account its variability due to the ultrasound image quality.

In complement, we counted the number of cases for which the EDNs produced results outside the inter-observer variability, *i.e.* a d_m value higher than 3.5 mm and 4.0 at ED and ES, respectively and a d_H value higher than 8.2 mm and 8.8 at ED and ES, respectively. From this experiment, we found that 18% of the segmentations produced by both U-Nets, ACNN or SHG can be seen as outliers. This value goes up to 30 % for U-Net++. For comparison purpose, the outliers rate from two series of annotations on fold 5 produced by the same expert O_{1a} is equal to 13%. Even if the overall performances of the EDNs are remarkable, this confirms the interest of still improving deep learning solution to produce highly reliable segmentation results on daily clinical practice data.

6) Accuracy of EDNs at Estimating Clinical Indices: Clinical scores provided in table IV show that EDNs produce results

below the inter-observer scores for all the metrics. It thus appears that the evaluated EDNs are serious candidates to automatically produce trustworthy estimations of the LV_{EDV} and LV_{ESV} indices, on par with medical expertise. Concerning the LV_{EF}, even if the results are better than the inter-observer scores, a correlation value of 0.82 (for the best performing method) appears too low in comparison with the intra-observer value of 0.90 to consider the automatic estimate of this index as sufficiently robust to be dependable in clinical practice. The lower LV_{EF} scores compared to LV_{EDV} and LV_{ESV} measures can be partially explained by the lack of temporal coherency in the tested EDN implementations. Indeed, for each patient, the ED and ES frames are viewed as two independent images, potentially generating less efficient estimation of the corresponding LV_{EF} measures. Numerous deep learning strategies that integrate temporal coherence such as the recurrent neural networks (the Long Short Term Memory - LSTM - model being one of the famous network of this family) has been described. The integration of such concepts into the U-Net formalism seems to be a solution of interest in order to make the LV_{EF} estimation more accurate.

VI. Conclusions

In this paper, we introduced the largest publicly-available and fully-annotated dataset for 2D echocardiographic assessment (to our knowledge). A dedicated Girder[2] on-line platform has been setup for new result submissions at https://camus.creatis.insa-lyon.fr/challenge/, where the CAMUS dataset, containing 2D apical four-chamber and two-chamber view sequences acquired from 500 patients, is also available for download. Thanks to this dataset, the following new insights were underlined:

- Encoder-decoder networks produced highly accurate segmentation results in 2D echocardiography;
- Among the different tested architectures, U-Net appeared to be most effective in terms of trade-off between the number of parameters and the achieved performance;
- The reasons for the lack of improvement of the more sophisticated networks (ACNN, SHG and U-Net++) compared to U-Net was addressed with additional testings provided in the supplementary materials/multimedia tab;
- U-Net reached a plateau in terms of its performances when training on more than 250 patients but still continued to improve, implying that though 250 patients was enough to generalize well on CAMUS, it has the potential to integrate additional variability;
- U-Net showed impressive robustness to variability, especially to image quality. Considering the wide range of image quality involved in echocardiography, this result is another positive element to consider encoder-decoder-based techniques as a solution of choice to solve the problem of 2D echocardiographic image segmentation;
- U-Net learned to reproduce a specific way of contouring;
- The segmentation and clinical results (left ventricle volumes at ED and ES) of the encoder-decoder networks were all below the inter-observer scores;
- The segmentation and clinical results of the encoder-decoder networks were close to but slightly worse than the intra-observer scores. This reveals that even if encoder-decoder networks produced remarkable results, there is still room for improvement to faithfully reproduce the manual annotation of a given expert.

[2]https://girder.readthedocs.io

References

[1] C. Armstrong *et al.*, "Quality control and reproducibility in M-mode, two-dimensional, and speckle tracking echocardiography acquisition and analysis: The CARDIA study, year 25 examination experience," *Echocardiography*, vol. 32, no. 8, pp. 1233–1240, 2015.

[2] O. Bernard *et al.*, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2514–2525, Nov. 2018.

[3] O. Bernard *et al.*, "Standardized evaluation system for left ventricular segmentation algorithms in 3D echocardiography," *IEEE Trans. Med. Imag.*, vol. 35, no. 4, pp. 967–977, Apr. 2016.

[4] D. Barbosa, D. Friboulet, J. D'hooge, and O. Bernard, "Fast tracking of the left ventricle using global anatomical affine optical flow and local recursive block matching," in *Proc. MICCAI Challenge Echocardiogr. Three-Dimensional Ultrasound Segmentation (CETUS)*, Boston, MA, USA, 2014, pp. 17–24.

[5] J. Pedrosa *et al.*, "Fast and fully automatic left ventricular segmentation and tracking in echocardiography using shape-based b-spline explicit active surfaces," *IEEE Trans. Med. Imag.*, vol. 36, no. 11, pp. 2287–2296, Nov. 2017.

[6] J. A. Noble and D. Boukerroui, "Ultrasound image segmentation: A survey," *IEEE Trans. Med. Imag.*, vol. 25, no. 8, pp. 987–1010, Aug. 2006.

[7] G. Carneiro, J. C. Nascimento, and A. Freitas, "The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 968–982, Mar. 2012.

[8] K. Y. E. Leung and J. G. Bosch, "Automated border detection in three-dimensional echocardiography: Principles and promises," *Eur. J. Echocardiogr.*, vol. 11, no. 2, pp. 97–108, 2010.

[9] C. Wang and Ö. Smedby, "Model-based left ventricle segmentation in 3D ultrasound using phase image," in *Proc. MICCAI Challenge Echocardiogr. Three-Dimensional Ultrasound Segmentation (CETUS)*, Boston, MA, USA, 2014, pp. 81–88.

[10] E. Smistad and F. Lindseth, "Real-time tracking of the left ventricle in 3D ultrasound using Kalman filter and mean value coordinates," in *Proc. MICCAI Challenge Echocardiogr. Three-Dimensional Ultrasound Segmentation (CETUS)*, Boston, MA, USA, 2014, pp. 65–72.

[11] M. Bernier, P. Jodoin, and A. Lalande, "Automatized evaluation of the left ventricular ejection fraction from echocardiographic images using graph cut," in *Proc. MICCAI Challenge Echocardiogr. Three-Dimensional Ultrasound Segmentation (CETUS)*, Boston, MA, USA, 2014, pp. 25–32.

[12] F. Milletari, M. Yigitsoy, N. Navab, and S. Ahmadi, "Left ventricle segmentation in cardiac ultrasound using Hough-forests with implicit shape and appearance priors," in *Proc. MICCAI Challenge Echocardiogr. Three-Dimensional Ultrasound Segmentation (CETUS)*, Boston, MA, USA, 2014, pp. 49–56.

[13] M. van Stralen, A. Haak, K. Leung, G. van Burken, and J. Bosch, "Segmentation of multi-center 3d left ventricular echocardiograms by active appearance models," in *Proc. MICCAI Challenge Echocardiogr. Three-Dimensional Ultrasound Segmentation (CETUS)*, Boston, MA, USA, 2014, pp. 73–80.

[14] O. Oktay, W. Shi, K. Keraudren, J. Caballero, and D. Rueckert, "Learning shape representations for multi-atlas endocardium segmentation in 3D echo images," in *Proc. MICCAI Challenge Echocardiogr. Three-Dimensional Ultrasound Segmentation (CETUS)*, Boston, MA, USA, 2014, pp. 57–64.

[15] K. Keraudren, O. Oktay, W. Shi, J. Hajnal, and D. Rueckert, "Endocardial 3D ultrasound segmentation using autocontext random forests," in *Proc. MICCAI Challenge Echocardiogr. Three-Dimensional Ultrasound Segmentation (CETUS)*, Boston, MA, USA, 2014, pp. 41–48.

[16] J. Domingos, R. Stebbing, and J. Noble, "Endocardial segmentation using structured random forests in 3D echocardiography," in *Proc. MICCAI Challenge Echocardiogr. Three-Dimensional Ultrasound Segmentation (CETUS)*, Boston, MA, USA, 2014, pp. 33–40.

[17] E. Smistad, A. Østvik, B. O. Haugen, and L. Løvstakken, "2D left ventricle segmentation using deep learning," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Sep. 2017, pp. 1–4.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[19] O. Oktay *et al.*, "Anatomically constrained neural networks (ACNNs): Application to cardiac image enhancement and segmentation," *IEEE Trans. Med. Imag.*, vol. 37, no. 2, pp. 384–395, Feb. 2018.

[20] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2016, pp. 424–432.

[21] E. D. Folland *et al.*, "Assessment of left ventricular ejection fraction and volumes by real-time, two-dimensional echocardiography. A comparison of cineangiographic and radionuclide techniques," *Circulation*, vol. 60, no. 4, pp. 760–766, 1979.

[22] R. M. Lang *et al.*, "Recommendations for cardiac chamber quantification by echocardiography in adults: An update from the american society of echocardiography and the European association of cardiovascular imaging," *J. Amer. Soc. Echocardiogr.*, vol. 28, no. 1, pp. 1–39, Aug. 2016.

[23] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[24] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Computer Vision—ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 483–499.

[25] Z. Zhou, M. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in in *Proc. Int. Workshop Deep Learn. Med. Image Anal. Int. Workshop Multimodal Learn. Clin. Decis. Support*, 2018, pp. 3–11.

[26] S. Leclerc, T. Grenier, F. Espinosa, and O. Bernard, "A fully automatic and multi-structural segmentation of the left ventricle and the myocardium on highly heterogeneous 2D echocardiographic data," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, 2017, pp. 1–4.

[27] P. Dollár and C. L. Zitnick, "Fast edge detection using structured forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1558–1570, Aug. 2015.

[28] D. Barbosa, T. Dietenbeck, J. Schaerer, and J. D'hooge, D. Friboulet, and O. Bernard, "B-spline explicit active surfaces: an efficient framework for real-time 3-D region-based segmentation," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 241–251, Jan. 2012.

[29] J. Pedrosa *et al.*, "Left ventricular myocardial segmentation in 3-D ultrasound recordings: Effect of different endocardial and epicardial coupling strategies," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 64, no. 3, pp. 525–536, Mar. 2017.

[30] D. Barbosa *et al.*, "Fast and fully automatic 3-D echocardiographic segmentation using B-spline explicit active surfaces: Feasibility study and validation in a clinical setting," *Ultrasound Med. Biol.*, vol. 39, no. 1, pp. 89–101, 2013.