LU-Net: A Multistage Attention Network to Improve the Robustness of Segmentation of Left Ventricular Structures in 2-D Echocardiography

Sarah Leclerc, Erik Smistad, Andreas Ostvik, Frédéric Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Mourad Belhamissi, Sardor Israilov, Thomas Grenier, et al.

LU-Net: a multi-stage attention network to improve the robustness of segmentation of left ventricular structures in 2D echocardiography

Sarah Leclerc, Erik Smistad, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Mourad Belhamissi, Sardor Israilov, Thomas Grenier, Carole Lartizien, Pierre-Marc Jodoin, Lasse Lovstakken, and Olivier Bernard

***Abstract*—Segmentation of cardiac structures is one of the fundamental steps to estimate volumetric indices of the heart. This step is still performed semi-automatically in clinical routine, and is thus prone to inter and intra observer variability. Recent studies have shown that deep learning has the potential to perform fully automatic segmentation. However, the current best solutions still suffer from a lack of robustness, in terms of accuracy and number of outliers. The goal of this work is to introduce a novel network designed to improve the overall segmentation accuracy of left ventricular structures (endocardial and epicardial borders) while enhancing the estimation of the corresponding clinical indices and reducing the number of outliers. This network is based on a multi-stage framework where both the localization and segmentation steps are optimized jointly through an end-to-end scheme. Results obtained on a large open access dataset show that our method outperforms the current best performing deep learning solution with a lighter architecture and achieved an overall segmentation accuracy lower than the intra observer variability for the epicardial border (*i.e.* on average a mean absolute error of 1.5 mm and a Hausdorff distance of 5.1 mm) with 11% of outliers. Moreover, we demonstrate that our method can closely reproduce the expert analysis for the end-diastolic and end-systolic left ventricular volumes, with a mean correlation of 0.96 and a mean absolute error of 7.6 ml. Concerning the ejection fraction of the left ventricle, results are more contrasted with a mean correlation coefficient of 0.83 and an absolute mean error of 5.0%, producing scores that are slightly below the intra observer margin. Based on this observation, areas for improvement are suggested.**

***Index Terms*—Cardiac segmentation, cardiac diagnosis, localization, deep learning, ultrasound, left ventricle, myocardium**

I. Introduction

Analysis of 2D echocardiographic images based on the measurement of cardiac morphology and function is essential for diagnosis. Low-level image processing such as segmentation and tracking enable to extract and interpret clinical indices, among which the volume of the left ventricle (LV) and the corresponding ejection fraction (LV_{EF}) are among the most commonly used. The extraction of such measures requires accurate delineation of the left ventricular endocardium (LV_{Endo}) at both end diastole (ED) and end systole (ES). However, these indices are subject to controversy due to a lack of reproducibility. Indeed, there is a significant variability in the measurement of the values extracted from the ultrasound images from an inter-expert, intra-expert and inter-equipment perspective. 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 the 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 across patients and pathologies.

S. Leclerc, T. Grenier, S. Israilov, M. Belhamissi C. Lartizien, F. Cervenansky and O. Bernard are with the Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1206, F-69621, LYON, France. E-mail: 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, Trondheim, Norway

F. Espinosa is with the Cardiovascular department Centre Hospitalier de Saint-Etienne 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, Trondheim, Norway

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

A. Related works

Numerous studies have been conducted for more than 30 years to make automatic measurements of the LV_{Endo} and LV_{EF} indices robust and reliable in echocardiographic imaging. Traditional methods correspond to deformable models [1], [2], motion-based methods [3], graph-based approaches [4], active appearance models [5], atlas-based methods [6] and machine learning algorithms [7]–[9]. Most of these approaches rely on handcrafted features which may amount to an over-simplistic source of information.

Supervised deep learning methods rely on more flexible models that go beyond this limitation. By definition, these techniques are optimal to the data they are trained on. Such approaches have been applied in the context of left ventricular structures analysis [10], [11], in particular for segmentation. In 2012, Carneiro *et al.* exploited deep belief networks and the decoupling of rigid and nonrigid classifiers to improve robustness in terms of image conditions and shape variability [12]. Later, Chen *et al.* proposed to use transfer learning from cross domain to enhance feature representation [13]. Dong *et al.* developed a deep fusion network to achieve coarse segmentation of the LV on 3D echocardiography. The derived outcomes were then used to initialize a classical deformable model to further optimize the segmentation results [14]. In parallel, Smistad *et al.* showed that the U-Net method [15] could be trained with the output of a state-of-the-art deformable model to successfully segment the LV in 2D ultrasound images [16]. Oktay *et al.* further extended a U-Net model so that

its segmentation output was constrained to fit a non-linear compact representation of the underlying anatomy derived from an auto-encoder network [17]. Leclerc *et al.* showed that a simple U-Net model trained on a large annotated dataset produces accurate segmentation results that are much better than the state-of-the-art, on average lower than the inter observer variability and close but still above the intra observer variability with 18% outliers. Recently, two methods have been proposed to cope with multi-view echocardiographic sequences segmentation. These methods are based on a large scale dataset, composed of more than 10850 images (then updated to 13500 images) collected from 100 patients (then updated to 150 patients). For each patient, three types of acquisition were carried out (apical 4, 3 and 2 chamber views) and manual annotations were made over a complete cardiac cycle by several experts. Based on such a dataset, Li *et al.* first proposed a deep pyramid and deep supervision network to process each frame of the sequence independently [18]. Their model incorporates a densely connected network, a feature pyramid network and a deeply supervised network to extract and fuse multi-level and multi-scale semantic information. Interestingly, this method outperforms the baseline U-Net model for the segmentation of end-diastolic and end-systolic frames. Part of the same group then proposed another method based on a recurrent aggregation network in order to integrate temporal coherency during the segmentation of one full cardiac cycle [19]. While both of these methods are clearly promising, unfortunately neither the dataset nor the trained models have been made available by the authors, making any comparison of these methods extremely difficult.

Deep learning methods based on regression models without intermediate segmentation have also been studied to directly estimate clinical indices from cine-MR and echocardiographic sequences [20]–[22]. In this context, the work produced by Ge *et al.* is among the most advanced studies [22]. Based on a Res-circle network, their PV-LVNet model embeds both subject features and temporal changes to effectively perform localization, cropping and clinical indices regression. Although this type of method appears to give promising results, the fact that it does not produce segmentation outputs can be considered a weakness, since segmentation contours are largely used by cardiologists to visually control the quality of the computed clinical indices. We therefore decided to focus on the improvement of left ventricular segmentation methods in echocardiographic imaging.

B. Attention learning-based approaches

In parallel, there has been an increasing interest in the computer vision community for deep learning methods based on contextualization to improve classification [23], localization [24], [25] and segmentation tasks [26]. Specifically, these methods incorporate deep attention mechanisms to emphasize what part of a given set of features the network should focus on according to some learned weights [27]. Most of the time, this is done through an Hadamard product between a value tensor V and an attention tensor A. For soft attention, the attention weights are values between 0 and 1 typically computed with a softmax or a sigmoid [28], [29] while for hard attention weights are binary [30]. In that sense, hard attention comes down to applying a binary mask onto network values, which can be feature maps [29] or the input image [28], [31]. Thus, if the binary mask has a rectangular shape, the attention module ends up extracting (or cropping) a region of interest (ROI) from V.

Hard attention networks: one of the most famous methods in computer vision is the Mask R-CNN method recently proposed by He *et al.* [26]. This approach provides among the best results in all three tracks of the COCO suite of challenges. The corresponding network is composed of three stages: *i)* a region proposal network (RPN) which scans boxes distributed over the image area and finds the ones that contain objects; *ii)* a classification network that scans each of the regions of interest proposed by the RPN and assigns them to different classes while refining the location and size of the bounding box to encapsulate the objects; *iii)* a convolutional network that takes the regions selected by the ROIs classifier and generates masks (*i.e.* segmentations) for them. Note that the first two stages of this network correspond to the Faster R-CNN framework developed for object detection [32]. In echocardiography, mainly two approaches have used this concept either for regression or segmentation. Ge *et al.* deployed a hard-attention strategy before applying their regression network to estimate a set of clinical indices. In particular, a Res-circle network was applied to coherently detect the LV centers over the cardiac cycle. Based on this information, a cropping strategy with pre-defined fixed dimensions was then applied to generate new images centered on the LV cavity. Leclerc *et al.* also introduced a contextualization mechanism based on the multiplication of a binary map surrounding the union of the LV and the myocardium (derived from a first segmentation network) with the input image in order to provide as input a pre-processed image without irrelevant information to a U-Net model that performs the segmentation of left ventricular structures [33]. Results show that this method allows for a reduction of outliers in terms of segmentation results (from 20% to 16%) but unfortunately without any improvement in overall accuracy.

Soft attention networks: some of these techniques have been successfully applied in medical imaging [27], [31], [34], [35]. In particular, Schlemper *et al.* developed a generic attention model to automatically learn to focus on target structures in medical imaging [27]. Based on attention gate modules that can be integrated in any existing CNN architecture [34], the proposed formalism intrinsically promotes the suppression of irrelevant regions in an input image while highlighting salient features useful for a specific task. As far as we know, no soft-attention learning techniques have been applied so far for the segmentation of echocardiographic images.

C. Objectives

Based on the literature review carried out in Sec. I-A and I-B, we decided to investigate the capacity of attention-based networks to improve the current best segmentation scores obtained in 2D echocardiographic imaging. To the best of our knowledge this is the first time that such an evaluation

is performed. More specifically, the purpose of this paper is to provide answers to the following four questions:

1) What improvement can be brought by attention-based architectures compared to the current best deep learning methods for 2D echocardiographic segmentation ?
2) Can we adapt attention architectures to the specificities of echocardiographic images ?
3) Can the number of outliers be significantly reduced ?
4) Do attention-based networks produce results below the intra observer variability scores both in terms of segmentation and clinical index estimation ?

Since the CAMUS dataset [36] is the current largest open access 2D echocardiographic dataset, we decided to build our study on the corresponding data. In particular, this dataset is composed of two and four-chamber acquisitions of 2D echocardiographic sequences from 500 patients with reference measurements from one cardiologist on the full dataset and from three cardiologists on a fold of 50 patients. An evaluation platform is also maintained to easily compare the performance of proposed new methods. The different studies conducted so far on this dataset highlighted three interesting outputs: *i)* the U-Net model currently produced the best segmentation results; *ii)* the corresponding scores are not much sensitive to the choice of hyper-parameters which reinforces the quality of the results obtained by such architecture; *iii)* the use of more sophisticated encoder-decoder architectures (*i.e.* U-Net++ [37], stacked hourglasses network [38] and anatomically constrained neural network [17]) did not produce better results. Therefore, while U-Net appears as a good choice for the segmentation of echocardiographic images, the improvement of its performance through the extension of its architecture is not straightforward.

II. Methodology

A. Motivations

The work carried out in this study was motivated by an experiment we conducted on the CAMUS dataset, whose details are described below. In particular, we manually selected regions of interest (ROIs) around the reference segmentation masks. Each ROI corresponds to the ideal bounding box (BB) surrounding the corresponding mask with an additional margin m of 5, 15 and 30% along the axes. From these ROIs, the corresponding images were cropped to create new datasets that were processed with the baseline U-Net1 architecture described in [36]. The corresponding scores are reported in Table II and referred to as BB-m5, BB-m15 and BB-m30, respectively. From this table, it is worth noting the contribution of the cropping stage, leading to a significant improvement of the baseline U-Net1 results, with average scores all below the ones of the intra observer (except for BB-m30 with the Hausdorff distance metric) and a number of outliers lower than 8%. This experiment thus reveals that the effective insertion of a localization step during the segmentation process with the U-Net architecture would yield remarkable results in echocardiographic image segmentation.

B. Overall strategy

Based on the motivations and the literature review on attention learning presented in the previous sections, we developed a multi-stage network to improve the robustness of segmentation in 2D echocardiography. Since the U-Net model already produces high-performance segmentation results in echocardiography [36], we decided to use this architecture as backbone for our multi-stage network, referred to as Localization U-Net (LU-Net) in the sequel. LU-Net aims at locating the left ventricle before segmenting the endocardial and the epicardial borders through an end-to-end learning procedure. The underlying assumption of this strategy is that the joint optimization of these two tasks should lead to better segmentation results. An illustration of the LU-Net's overall architecture is provided in Fig. 1. In particular, LU-Net is composed of two main parts: one RPN for localization and one U-Net model for segmentation.

1) Backbone U-Net architecture: The same U-Net architecture was used in the localization and segmentation parts. It consists of an encoder and a decoder stage which have several layers of 3×3 2-D convolutional filters with ReLU activation functions. In the encoder stage, the input image was processed by an increasing number of filters followed by max pooling subsampling after the convolution layers. Reaching a final spatial size of 8×8, the decoder increases the spatial size gradually by upsampling and convolution stages with decreasing number of filters. In addition, the network has multiple skip connections between the encoder and the decoder to recover the fine-grained spatial details which may be lost after max pooling. Since the network was designed for real-time performances, we kept the number of layers and convolutions as low as possible and used 2D upsampling operations instead of transposed convolution for the decoder. The result is a network with about two million parameters which can do segmentation in a matter of milliseconds. Network input is a single image resized to 256×256 pixels, and the output is an image of the same size as the input with three channels. Each channel is a normalized logit for each class by softmax activation.

2) Localization network: The proposed RPN is composed of a combination of the U-Net model whose architecture is described above, followed by a standard regression network. The regression network consists of two main parts: *i)* a feature extractor identical to the downsampling part of the U-Net model described above, composed of 12 layers of 3×3 2-D convolutional filters with ReLU activation functions and *ii)* a multi-layer perceptrons (MLP) composed of 4 fully connected layers. The output of the network corresponds to the 4 relative coordinates of the bounding box (BB) around the structure of interest, namely the union of the left ventricle and myocardium, referred to as $(x_{min}, x_{max}, y_{min}, y_{max})$. The MLP connected to the flattened output of the feature extractor is composed of hidden layers of respectively 1024, 256 and 32 units and one final layer of 4 units without activation function in order to allow regression on the coordinates of the bounding box. The use of an initial segmentation as intermediate feature

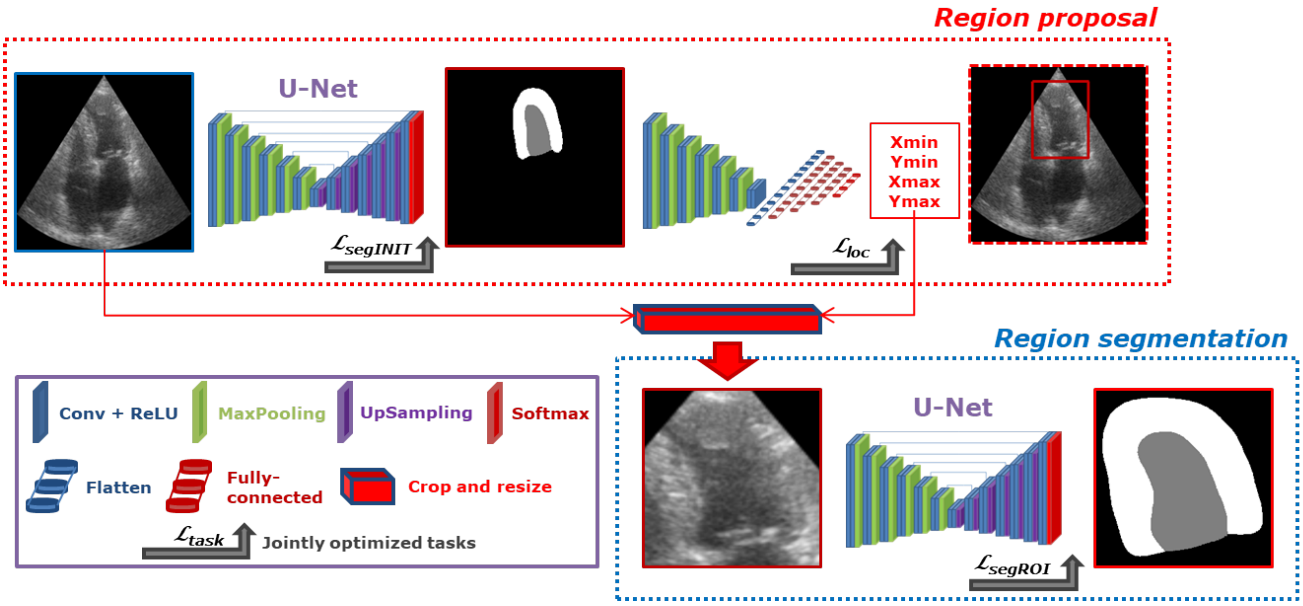

Fig. 1. Illustration of the LU-Net pipeline with the U-L2-mu localization network introduced in Sec. II-B2. The two U-Nets are independent.

maps allows us to benefit from the good overall performance of U-Net model to guide the regression network. Moreover, since the U-Net model we use was optimized in terms of number of parameters, our localization network presents the strong advantage to be lighter and faster than most state-of-the art approaches, as shown in Sec. III-C. While the reference BB were defined as the minimal bounding boxes in contact with the epicardium border, the target coordinates were computed with an additional margin m as:

$$x^m_{min} = x_{min} - m * h \quad , \quad x^m_{max} = x_{max} + m * h,$$
$$y^m_{min} = y_{min} - m * w \quad , \quad y^m_{max} = y_{max} + m * w.$$

where (w, h) are the width and height of the reference BB. The motivation for adding a margin was to provide some context around the targeted structures for the segmentation task.

3) Segmentation network: The output of the region proposal network is used as an attention mechanism to crop and resize the input ultrasound image. The resulting image is fed to a second segmentation network whose architecture corresponds to the one of the U-Net model described in Sec. II-B1. It is worth noting that this model is currently the most efficient one evaluated on the CAMUS dataset considering a trade-off between accuracy, speed and size [36].

4) End-to-end approach: In order to make the full network trainable end-to-end, the cropping of the input image and the resizing of the corresponding ROI (red block in Fig. 1) are realized by bilinear interpolation using only differentiable operations[1]. During the computation of the segmentation loss involved in the second U-Net, it is also necessary to crop the ground truth mask according to the predicted BB so that the same regions are taken into account in the calculation. This makes the second segmentation loss function evolve dynamically over the training phase. This step is realized using the same bilinear differentiable sampling strategy as the one mentioned above. The two U-Nets involved in our architecture are distinct networks whose weights are learned simultaneously. Three main sequential tasks are thus trained at the same time during the optimization process: *i)* the initial U-Net segmentation used in our RPN; *ii)* the localization of the LV bounding box using the initial segmentation; *iii)* the final U-Net segmentation performed on the ultrasound image cropped using the bounding box from the localization network. Taken together, the different aspects described in this section allows the gradients to flow all the way from the output to the input of the network. At inference time, based on the localization outputs, the final segmentation result is then returned to the original coordinate system of the input image.

[1]In practice, this step is carried out through the function crop_and_resize of the TensorFlow library

III. Experiments

A. Dataset

The CAMUS dataset contains two and four-chamber acquisitions from 500 patients [36]. The full dataset was divided into 10 folds equally distributed in terms of image quality (good, medium, poor) and ejection fraction category ($\leq$ 45%, $\geq$ 55% or in between). This allows the analysis of the full dataset by means of a classical cross-validation strategy. One cardiologist (O_1) manually annotated the endocardium and epicardium (LV_{Epi}) borders of the left ventricle on the full dataset at end diastole (ED) and end systole (ES) and two other cardiologists (O_2 and O_3) on a fold of 50 patients. This fold was also annotated twice by O_1 seven months apart. This procedure allows comparison of the results provided by the algorithms with the inter- and intra observer variability. Since the work provided in [36] concluded that the current best solutions produced results all below the inter-observer scores but still worse than the intra observer ones, we focused in

this article on comparing the results obtained by the different evaluated methods with the intra observer variability.

B. Evaluation metrics

1) Localization metrics: We assessed the performance of the localization networks through the Intersection Over Union (IOU) metric and the euclidean distance errors between the predicted and the reference BB coordinates (*i.e.* its central position (x_c, y_c), its height h and width w). The IOU is a classical localization metric which measures the overlap between the predicted BB and the reference one. It gives a value between 0 (no overlap) and 1 (full overlap). In addition, we provided the "BB out" metric which corresponds to the number of cases where the predicted BB does not completely encompass the reference mask.

2) Segmentation metrics: To measure the accuracy of the segmentation output (LV_{Endo} and LV_{Epi}) of a given method, the Dice metric (closely related to the IOU and classically used in segmentation), the mean absolute distance (d_m) and the 2D Hausdorff distance (d_H) were used. The Dice similarity index is a measure of overlap between the segmented surface S_{user} extracted from a method and the corresponding reference surface S_{ref}. It 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 maximum local distance between the two surfaces. In addition, we assessed the quality of segmentation with regard to cardiologists' annotations through the notion of outliers defined below.

- ***Geometric outlier***: the set of segmentation attached to a patient is seen as a geometric outlier if at least one of its eight corresponding distance scores (*i.e.* d_m and d_H values at ED and ES for both apical two and four-chamber views) is out of the corresponding bounds defined from the inter-observer variability [36];
- ***Anatomical outlier***: the set of segmentation attached to a patient is seen as an anatomical outlier if the simplicity and convexity [33] of the corresponding segmented contours are lower than the lowest values computed from expert annotations on 50 patients. These two metrics hold values between 0 and 1, and are maximized for a circle. They also give discriminating values for any convex shapes, such as oval shapes like heart cavities, and bridge shapes like the myocardium. They can therefore be used as simple tools to detect anatomical outliers in the case of left ventricular structures.

3) Clinical metrics: We evaluated the performance of the methods with 3 clinical indices: *i)* the ED volume (LV_{EDV} in ml); *ii)* the ES volume (LV_{ESV} in ml); *iii)* the ejection fraction (LV_{EF} as a percentage), for which we computed two metrics: the correlation ($corr$) and the limit of agreement (loa) ($mean$ ± 1.96 std). All left ventricular volumes were computed using Simpson's biplane rule [39], involving the segmentation results on both two- and four-chamber apical views.

C. Localization methods

We implemented and assessed the performance of five convolutional networks dedicated to the prediction of bounding boxes, *i.e.* predicting $(x^m_{min}, x^m_{max}, y^m_{min}, y^m_{max})$.

1) An AlexNet-like network composed of two parts: a feature extractor corresponding to the original Alexnet architecture (details on the corresponding network can be found in [40]) whose flattened output is connected to a MLP with two hidden layers of 4096 units each and a final layer of 4 units. No dropout nor data augmentation was performed. This model has 71M parameters.
2) A VGG19-like network composed of two parts: a feature extractor corresponding to the original VGG19 architecture (details on the corresponding network can be found in [41]) whose flattened output is connected to a MLP with two hidden layers of 4096 units each and one final layer of 4 units. This model has 70M parameters.
3) A Faster R-CNN model [24] composed of three stages: *i)* a feature pyramid network to extract features using the Resnet101 architecture as backbone [42], *ii)* a region proposal network based on anchors that scans the image in a sliding-window fashion to find areas that contain objects of interest; *iii)* a ROI classifier, associated to a bounding box regressor to further refine the location and size of the bounding boxes that encapsulate the recognized objects. The network has 61M parameters.
4) A U-L1 model based on the U-Net architecture described in Sec. II-B1 to perform the segmentation of the left ventricle and the myocardium. The bottom layer of this U-Net was derived in order to carry out the localization procedure using four fully connected layers of 1024, 256, 32, and 4 units. This model was inspired by the work of Vigneault *et al.* [31]. This network has 9M parameters;
5) A U-L2 model also based on the U-Net architecture described in Sec. II-B1 to perform the segmentation of the left ventricle and the myocardium. The output of this U-Net was then connected to a downsampling branch ending with four fully-connected layers of 1024, 256, 32 and 4 units. More details on this branch can be found in Sec. II-B2. Contrary to most of the localization networks found in the literature, this model uses a pre-segmentation network to guide a localization procedure. We evaluated two versions of this network, one optimizing only the localization loss (referred to as U-L2-mo) and one optimizing both the localization and the segmentation losses (referred to as U-L2-mu). The network includes 11M parameters.

D. Segmentation methods

The performance of the joint segmentation of the endocardial and the epicardial borders was assessed through the following five networks:

1) U-Net1, corresponding to the current best performing network on the CAMUS dataset [36]. This network includes 2M parameters.

2) RU-Net, recently introduced in [33] and built from two cascaded U-Net1. The epicardial mask predicted by the first network is dilated and multiplied with the input image to provide a contextualized image as input to the second network, with a total number of 4M parameters.
3) Attention-gated U-Net (AG-U-Net), recently proposed in [34], in which attention layers are used at each skip connection to locally weigh the concatenated features with coefficients derived from the previous layer. It includes batch normalization before each activation, and deep supervision by aggregating the feature maps produced after each attention layer at the last level of U-Net1 (*i.e.* before the last convolution and the softmax). This network has a total number of 2M parameters.
4) Mask R-CNN, recently proposed in [26] and built upon the Faster R-CNN model described in Sec. III-C. More specifically, this network consists of a RPN based on the Faster R-CNN architecture followed by a segmentation network composed of five convolution layers producing low resolution masks (28×28 pixels) that are scaled up to the size of the ROI bounding box at inference time. This network includes 64M parameters.
5) LU-Net, as introduced in this paper, built using U-L2-mu as the region proposal network and U-Net1 as the segmentation network for a total of 13M parameters. Two margins of $m = 5\%$ and $m = 15\%$ were evaluated.

E. Learning strategy

1) Loss: Localization networks were optimized using a L1 loss clipped at 0.99 summing the errors on the four relative BB values (*i.e.* $(x^m_{min}, x^m_{max}, y^m_{min}, y^m_{max})$). Segmentation networks were optimized using a multi-class Dice loss (inspired from [43]) taking into account the LV and myocardium predictions. The overall loss of LU-Net is given by:

$$\mathcal{L} = \mathcal{L}_{seg_{INIT}} + \lambda * \mathcal{L}_{loc} + \mathcal{L}_{seg_{ROI}} \quad (1)$$

with:

$$\mathcal{L}_{seg_{INIT}} = 1 - 2\frac{\sum_{l=0}^{2}\sum_n p_{ln} gt_{ln}}{\sum_{l=0}^{2}\sum_n p_{ln} + gt_{ln}} \quad (2)$$

$$\mathcal{L}_{loc} = \sum_{i-1}^{4} |BB_i^m - BB_i^{gt}| \quad (3)$$

$$\mathcal{L}_{seg_{ROI}} = 1 - 2\frac{\sum_{l=0}^{2}\sum_n p_{ln[ROI]} gt_{ln[ROI]}}{\sum_{l=0}^{2}\sum_n p_{ln[ROI]} + gt_{ln[ROI]}} \quad (4)$$

In the above equations, BB_i^{gt} stands for the i-th ground truth bounding box coordinates. $p_{ln} = 1$ if the predicted segmentation map p has label l over pixel n (the same applies to the ground truth segmentation map gt). λ is a fix coefficient that balances the localization and segmentation losses.

2) Parameter settings: All the methods involved in this study were optimized using the Adam optimizer associated to a learning rate (either equal to $1e^{-3}$ or $1e^{-4}$) and a number of epochs (controlled using early stopping with the patience parameter set to 20) that experimentally allowed to observe a smooth convergence of the training and validation losses. In particular, we set a maximal number of epochs equal to 100 for the LU-Net method. The best model on the validation loss was selected after each training phase. Several experiments have been carried out in order to determine the optimal value of the λ coefficient that balances the localization and segmentation losses (see Eq. 1). Experimentally, we found that a value of 10 produced the best results. However, we also noted that this parameter was not sensitive, and that values that deviate from the optimal value produce comparable results.

IV. Results

In order to easily compare our results with those of the state-of-the-art on the CAMUS dataset, we followed the strategy developed in [36] by training for each deep learning method a single model on the annotated images of both apical two and four-chamber views, regardless of the time instant.

A. Localization results

Table I shows the localization accuracy computed on the full dataset (500 patients) for the five algorithms described in section Sec. III-C. Mean and standard deviation values for each metric were obtained from cross-validation on the 10 folds of the dataset (see [36] for more details). For each row of this table, the *m* information indicated after the name of the method indicates the margin value used to define the reference BB. As for Faster R-CNN, we stayed as close as possible to the original implementation [26] in order not to bias the results. This model is designed so that the predicted bounding box encompasses the object of interest as closely as possible. This is the reason why we do not mention any *m* information for this method. The values in bold displayed in Table I correspond to the best scores for each metric.

As one can see, Faster R-CNN method produces the best scores in terms of IOU value and errors on the estimation of the coordinates of the bounding box compared to methods involving a margin. This is probably due to the fact that the addition of a margin makes it more difficult to accurately locate the object of interest since the edges of the bounding box no longer correspond to the borders of the object, which generally have sharp edges that are easier to detect. However, the absence of a margin drastically increases the number of cases where the reference mask is not fully encompassed by the estimated bounding box, with a BB out value equal to 90% of the cases. Finally, as the Faster R-CNN model only detects bounding boxes associated with a sufficiently high level of confidence, there is no guarantee that this model will detect BB for each processed image. During our experiments, 53 cases over 1624 (3%) of test images were missed. Theses cases were excluded when we computed the Faster R-CNN scores given in Table I.

Based on a comparison of the methods using a margin of 5%, the proposed U-L2-mu gets the overall best localization scores on all metrics, except for the error on y_c with a difference of 0.2 mm with the best method. These results validate the use of the U-Net architecture, which has already proven its effectiveness in terms of segmentation, to perform

TABLE I
LOCALIZATION ACCURACY OF THE FIVE METHODS DESCRIBED IN SEC. III-C AND EVALUATED ON THE FULL DATASET (500 PATIENTS). THE m INFORMATION CONTAINED IN EACH METHOD NAME INDICATES THE MARGIN VALUE DEFINED IN SEC. II-B2

Model	**IOU**	**Error (mm)**				**BB out**
		x_c	y_c	h	w	
Faster R-CNN	**0.909** ±0.042	**1.3** ±1.2	**1.3** ±1.1	**2.9** ±2.6	**2.8** ±2.6	1797 90%
AlexNet-m5	0.880 ±0.062	2.2 ±2.4	1.9 ±1.8	4.2 ±4.1	4.1 ±4.1	866 43%
VGG-m5	0.888 ±0.060	1.9 ±2.4	1.7 ±1.7	4.0 ±3.9	4.0 ±3.7	903 45%
U-L1-m5	0.849 ±0.072	3.1 ±2.9	2.7 ±2.4	5.3 ±4.5	4.9 ±4.3	1094 55%
U-L2-mo-m5	0.791 ±0.138	4.2 ±4.7	4.4 ±6.0	7.1 ±6.4	6.9 ±6.7	1393 70%
U-L2-mu-m5	0.898 ±0.053	1.6 ±1.8	1.9 ±1.9	3.2 ±3.1	3.6 ±3.2	712 36%
U-L2-mu-m15	0.907 ±0.054	1.6 ±2.0	1.7 ±1.7	3.7 ±4.0	4.3 ±4.3	31 2%

localization tasks in ultrasound imaging compared to well-established computer vision architectures (*i.e.* AlexNet and VGG). In addition, the scores highlight the interest of using both segmentation and localization losses to improve the performance of the U-L2 method, with an average gain of 2.5 mm over the BB centre estimate and 3.6 mm over the BB dimension estimate. This significant improvement demonstrates that forcing segmentation as an intermediate step to localization is beneficial.

We also investigated the influence of the choice of the margin value *m* on the accuracy of the localization results produced by the U-L2 method. The obtained results are contrasted. Indeed, while the use of a lower margin (*i.e.* 5%) produces slightly better results with regard to the estimation of the BB position, the use of a higher margin (*i.e.* 15%) considerably reduces the number of cases where the BB does not encompass the reference mask (from 36% to 2%). Based on this experiment, it is clear that the U-L2-mu model produced among the best localization results with the lightest architecture. We therefore decided to use this network as the region proposal part of the LU-Net architecture, as illustrated in Fig. 1. Moreover, while a large value of *m* would ensure to encompass the left ventricular region for all cases, it would be at the cost of the contextualization effect we are looking for. There is therefore a trade-off between the value of the margin used for the localization network and the accuracy of the final stage segmentation network. This is the reason why we assessed the segmentation accuracy of our LU-Net model for two different margin values (5% and 15 %) in the rest of the experiments.

B. Segmentation results

Table II displays the segmentation accuracy computed on the full dataset from patients having good and medium image quality (406 patients) for the five algorithms described in section Sec. III-D. Mean and standard deviation values for each metric were obtained from cross-validation 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 Mask R-CNN produces competitive results compared to the baseline U-Net1, with better results for the Hausdorff distances. However, for 3% of the cases this network failed to detect a bounding box and therefore failed to produce segmentation results. These 3% of cases were not taken into account when calculating the segmentation scores, but were counted as outliers. As for the other attention networks, they produced either the same, or better results than the baseline U-Net1, with AG-U-Net and LU-Net being the best performing models. Indeed, AG-U-Net obtained the overall best results for the segmentation of the LV_{Endo} border (d_m value of 1.5 mm and d_H value of 5.3 mm), leading to segmentation scores close but still higher than the intra observer variability for this structure. The LU-Net-m5 approach obtained the best results for the segmentation of the LV_{Epi} border (d_m value of 1.5 mm and d_H value of 5.1 mm) and the lowest number of geometric outliers (11%). Interestingly, these scores are either equivalent or lower than the intra observer variability for this structure. It is also worth noting the robustness of the LU-Net model with respect to the choice of margin parameter, as margins of $m = 5\%$ and $m = 15\%$ produce almost the same segmentation scores for all metrics. An illustration of the segmentation performance of the LU-Net-m5 network compared to the baseline U-Net1 model on three different cases is provided in Fig. 2.

C. Clinical scores

Table III contains the clinical metrics computed on the full dataset from patients having good and medium image quality (406 patients) for the five methods described in Sec. III-D. Those indices were computed with the Simpson's biplane rule [39] from the segmentation results of each algorithm on the two- and four-chamber apical views. The values in bold represent the best scores for the corresponding index. For the Mask R-CNN method, 12% of patients (36 cases) did not have a prediction for all necessary four images to compute the EF and were therefore not included in the presented clinical results. While the corresponding scores are competitive for LV_{EDV} and LV_{EF}, it appears that this model has a strong tendency to underestimate the LV_{EDV}, with a bias between 6 to 10 times higher that the other methods and a *mae* score of 12.8 ml. As for segmentation, the AG-U-Net and LU-Net-m5 models 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}, the two methods produced high correlation scores (0.956), small biases (± 1.4 ml) and reasonable limit of agreements (around 22 ml) and mean absolute errors (around 8.3 ml). The AG-U-Net produced the best LV_{ESV} results with a correlation of 0.962, while the LU-Net-m5 model produced the best LV_{EF} scores with a correlation of 0.829. However, even if the scores of LU-Net-m5 and AG-U-Net are slightly better than the baseline U-Net1 ones, they are still higher than the intra observer results. This reveals that there is still room for improvement as discussed in Sec. V.

TABLE II
SEGMENTATION ACCURACY OF THE FIVE METHODS DESCRIBED IN SEC. III-D AND EVALUATED ON PATIENTS HAVING GOOD AND MEDIUM IMAGE QUALITY (406 IN TOTAL). THE m INFORMATION CONTAINED IN EACH METHODS NAME INDICATES THE MARGIN VALUE DEFINED IN SEC. II-B2

	Model	LV_{Endo}			LV_{Epi}			outliers
		D	***d_m***	***d_H***	***D***	***d_m***	***d_H***	**geo.**
		val.	mm	mm	val.	mm	mm	# %
	intra observer	0.937 ±0.027	1.4 ±0.5	4.5 ±1.8	0.954 ±0.020	1.7 ±0.8	5.0 ±2.2	21 13%
Motivation study (Sec. II-A)	BB-m5	0.941 ±0.034	1.3 ±0.6	4.3 ±1.9	0.971 ±0.011	1.0 ±0.4	4.1 ±1.8	89 5%
	BB-m15	0.940 ±0.034	1.3 ±0.6	4.4 ±1.9	0.969 ±0.011	1.1 ±0.4	4.3 ±2.0	106 6%
	BB-m30	0.937 ±0.035	1.4 ±0.6	4.7 ±2.1	0.966 ±0.013	1.2 ±0.5	4.6 ±2.2	124 8%
Experimental study (Sec. IV-B)	Mask R-CNN [26]	0.924 ±0.038	1.7 ±0.8	5.2 ±2.7	0.946 ±0.023	1.9 ±0.9	5.7 ±2.6	283 17%
	U-Net1 [36]	0.920 ±0.056	1.7 ±1.2	5.6 ±3.3	0.947 ±0.030	1.9 ±1.1	6.2 ±3.7	282 17%
	RU-Net [33]	0.925 ±0.049	1.7 ±1.0	5.4 ±3.3	**0.950** ±0.030	1.8 ±1.1	5.8 ±3.9	240 15%
	AG-U-Net [34]	0.930 ±0.049	**1.5** ±1.3	**5.3** ±3.4	**0.950** ±0.026	1.8 ±1.0	5.9 ±3.7	270 17%
	LU-Net-m5	**0.953** ±0.026	1.7 ±0.9	5.5 ±3.6	0.932 ±0.043	**1.5** ±0.8	**5.1** ±3.3	**186** 11%
	LU-Net-m15	0.952 ±0.029	1.7 ±1.1	5.6 ±4.0	0.931 ±0.049	**1.5** ±1.1	5.3 ±3.6	203 12%

* *LV_{Endo}: Endocardial contour of the left ventricle; LV_{Epi}: Epicardial contour of the left ventricle*
D: Dice index; d_m: mean absolute distance; d_H: Hausdorff distance
The values in bold refer to the best performance for each measure.

TABLE III
CLINICAL METRICS OF THE 5 EVALUATED METHODS DESCRIBED IN SEC. III-D AND RESTRICTED TO PATIENTS HAVING GOOD AND MEDIUM IMAGE QUALITY (406 IN TOTAL)

Model	LV_{EDV}			LV_{ESV}			LV_{EF}		
	corr	***loa***	***mae***	***corr***	***loa***	***mae***	***corr***	***loa***	***mae***
	val.	ml	ml	val.	ml	ml	val.	%	%
intra observer	0.978	-2.8±14.3	6.2	0.981	-0.1±11.4	4.5	0.896	-2.3±11.2	4.5
U-Net1 [36]	0.947	-8.3±24.7	10.9	0.955	-4.9±19.4	8.2	0.791	-0.5±15.1	5.6
RU-Net [33]	0.946	-1.2±23.9	8.9	0.949	0.3±19.6	7.3	0.704	-2.1±14.3	6.0
AG-U-Net [34]	**0.956**	-1.4±21.9	**8.1**	**0.962**	0.6±17.0	**6.2**	0.798	-2.2±15.1	5.5
Mask R-CNN [26]	0.953	-11.0± 24.9	12.8	0.955	-4.7±20.8	7.9	0.817	-2.6±13.9	5.8
LU-Net-m5	**0.956**	1.4 ±21.8	8.3	0.956	1.6± 18.0	7.0	**0.829**	-1.5±13.5	**5.0**
LU-Net-m15	0.952	2.4 ±22.9	**8.1**	**0.962**	1.8±16.7	6.5	0.821	-1.2±13.7	**5.0**

* *corr: Pearson correlation coefficient; loa: limit of agreement; mae: mean absolute error.*
The values in bold refer to the best performance for each measure.

D. LU-Net behavior

From the results given in Table II and Table III, it appears that the LU-Net method outperforms the baseline U-Net1 model both in terms of segmentation and clinical indice estimation. Furthermore, it is one of the most effective model, even compared to other attention-based networks. In order to complete the analysis of LU-Net, we applied this network to the full dataset (including poor image quality) and studied the generated outliers. The corresponding results obtained with a margin of $m = 5\%$ are provided in Table IV. The results of the model named LU-Net-m5-o1 corresponds to the scores derived from the output of the first U-Net involved in the region proposal network, while the scores of the model named LU-Net-m5-o2 corresponds to the scores derived from the final output of the network (*i.e.* the one provided by the second U-Net). From this table, one can see that LU-Net outperforms the U-Net1 architecture for all the metrics for both LV_{Endo} and LV_{Epi} borders when considering all quality of images. Also, the segmentation results produced by LU-Net appear to be remarkably stable when integrating poor image quality images, with a mean difference of 0.1 mm for d_m, 0.2 mm for d_H and 1% for the geometric outliers.

Concerning the localization scores, the LU-Net-m5 model obtained consistent results with respect to the U-L2-mu best performing method (among the implementations involving

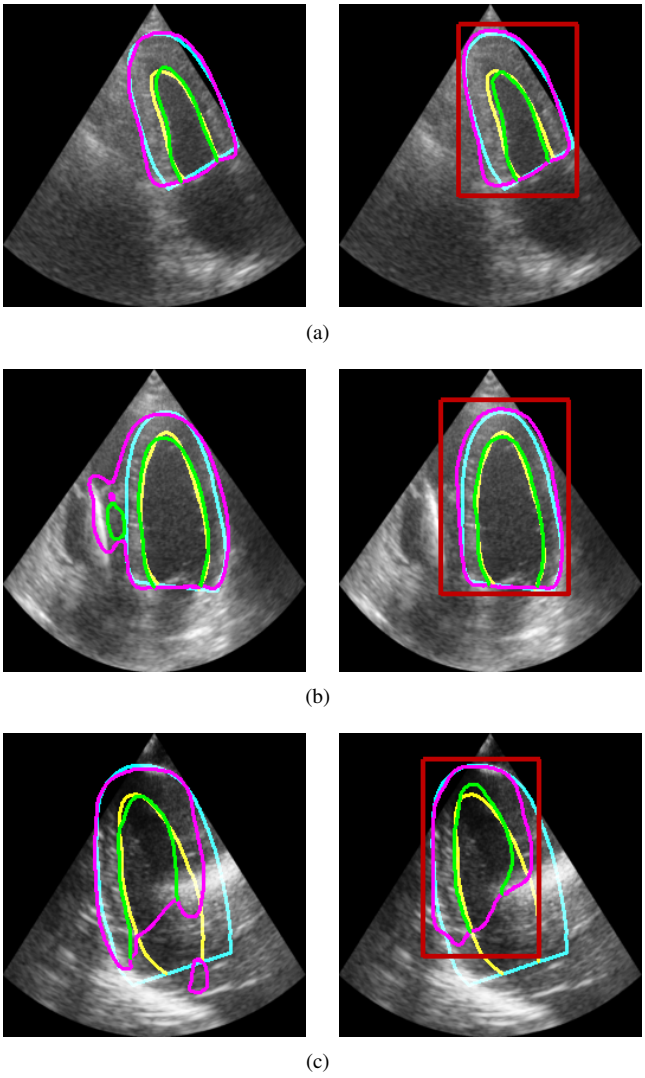

Fig. 2. Comparison of the segmentation performance of the baseline U-Net1 (left column) and the proposed LU-Net architecture (right column) on cases (a) with similar results; (b) where the intermediate localization of the LU-Net helps; (c) where the artifact present in the image is too strong for any improvement. In each image, the prediction is in green and purple while the ground-truth is in yellow and cyan. The BB estimated is displayed in red.

a margin) reported in Table I with an IOU of 0.906 and (x_c, y_c, h, w) BB errors of $(1.5, 1.5, 3.3, 3.5)$ mm, respectively. Coupling this result with the last two lines of Table IV which show that the first segmentation is less accurate than a single U-Net1, it appears that the segmentation result produced in the region proposal part of the LU-Net is degraded by optimizing the localization procedure, which in turn allows for a significant improvement of the final segmentation results compared to the baseline U-Net1 model.

Concerning the segmentation scores, LU-Net-m5 produced 12% of geometric outliers, 2% of anatomical outliers and 1% of both, showing that half of the anatomical outliers are also geometric. Moreover, the geometric outlier rate is lower than the intra observer variability one computed from a subset of 40 patients with good and medium image quality, which further highlights the quality of the results achieved by LU-Net.

TABLE IV
SEGMENTATION ACCURACY AND OUTLIERS ON THE FULL DATASET (500 PATIENTS) INCLUDING THOSE WITH POOR IMAGE QUALITY

Model	LV_{Endo}		LV_{Epi}		outliers		
	d_m	d_H	d_m	d_H	geo.	ana.	both
	mm	mm	mm	mm	# %		
U-Net1	2.0 ±1.2	6.1 ±3.9	2.0 ±1.1	6.5 ±4.5	423 21%	95 5%	71 4%
LU-Net-m5-o1	2.1 ±1.1	7.0 ±4.7	1.9 ±1.0	6.2 ±3.4	483 24%	201 10%	138 7%
LU-Net-m5-o2	**1.8** ±1.0	**5.7** ±3.6	**1.6** ±0.9	**5.3** ±3.3	**240** 12%	**31** 2%	**20** 1%

V. DISCUSSION

A. LU-Net versus Mask R-CNN

In this study, we compared the performance of LU-Net with Mask R-CNN [26], one of the most popular networks for joint localization and segmentation tasks. Although both methods are based on a localization step followed by a segmentation task, their architectures are very different. While our localization network uses a combination of a U-Net model with a simple regression network, the localization network of Mask R-CNN is based on three complex stages described in Sec. III-C. As far as segmentation goes, our model exploits a second U-Net while Mask R-CNN uses a simple set of five convolution layers producing low resolution masks (28×28 pixels) that are scaled up to the size of the ROI bounding box at inference time. Because of a more complex strategy, Mask R-CNN comprises 64M parameters, while our lighter model has 13M parameters, making it more suitable for the real time nature of ultrasound image processing. In terms of results, while the Faster R-CNN part of the Mask-R-CNN produces better localization results, LU-Net produces better segmentation scores for all metrics (apart for d_H concerning LV_{Endo}) as well as better estimation of all tested clinical indices. This confirms our initial motivation to provide as input to a U-Net model a cropped and resized region that fully encompasses the union of the left ventricle and myocardium with a relatively small margin value.

B. Attention-based networks

Table II and Table III underline the ability of attention-based networks to improve the segmentation and the estimation of clinical indexes in 2D echocardiography. These results are even more interesting given that the authors of the original study [36] had not succeeded in improving the scores of the baseline U-Net1 model through more sophisticated architectures. Although AG-U-Net produced the best scores on the LV_{Endo} and the estimation of the LV_{ESV}, LU-Net provides the best trade-off between the achieved improvements and the decrease of the number of geometric outliers.

C. Comparison with intra observer variability

As for the segmentation scores, the LU-Net model manages to reach the intra observer variability for the LV_{Epi} border (d_m

and d_H metrics). The number of geometric outliers, 11%, is also reduced below the intra observer rate. To the best of our knowledge, this is the first time that such result is obtained in the context of 2D echocardiographic image segmentation. In addition, one can observe that the scores reached by our model are still slightly higher than the intra observer variability for the LV_{Endo} border. Concerning the estimation of the clinical metrics, although LU-Net improves the results compared to the baseline U-Net1 model, its scores are still slightly higher than the intra observer variability. This reveals that while attention-based networks clearly enhanced the results produced by the baseline U-Net1 model, there still exists room for improvement to faithfully reproduce the manual annotations of one expert.

D. Areas for improvement

We identified two leads of potential improvement to allow competitive results with respect to the intra observer variability. First, based on Table I, it appears that the localization step can be further optimized to improve the LU-Net scores, as suggested by the results on ideal cases provided in Table II. Secondly, there is a need to introduce temporal coherency into deep learning architectures. Indeed, while the current strategy (*i.e.* ED and ES are treated independently) provides high correlations for the estimation of the LV_{EDV} and LV_{ESV} (0.956 for both indices), the estimation of the LV_{EF} is degraded to 0.829. This reveals the lack of temporal consistency of the LU-Net segmentation results between ED and ES.

E. Industrial applications

From a pure application stand-point, LU-Net does not rely on a recurrent network like an RNN or LSTM which is a great advantage. As such, our method does not need the entire cardiac cycle to segment the heart as it works on an image-by-image basis. This is inline with clinical needs where cardiologists can see in real time the segmentation of the heart as they weep their probe. In addition, what makes our method attractive for industrial applications is the fact that it is built around a U-Net architecture optimized according to the performance/speed ratio which can perform accurate segmentation in milliseconds, as proven in [36]. Since both the localization and segmentation branches use this architecture, it results in an overall architecture composed of 13M of parameters. This makes the inference run-time of LU-Net an average of $0.18s \pm 0.05s$ per frame, hence allowing near real time multi-structure segmentation. Thus, some efforts still need to be made to slightly reduce this execution time for direct use in industrial applications.

VI. Conclusions

In this paper, we introduced a novel multi-stage attention network to improve the robustness of segmentation of left ventricular structures in 2D echocardiography. Our network is built around the U-Net architecture and is composed of two stages: a region proposal network and a segmentation network. The performance of our solution was assessed on the current largest open access 2D echocardiographic dataset. Using this collection of data, the following contributions have been achieved:

- A review of five different localization methods (including well-established networks applied in computer vision and innovative architectures adapted to echocardiography) was performed to assess the accuracy of localization of LV structures in 2D echocardiography.
- A novel localization architecture was proposed. In particular, we showed that the combination of a U-Net to pre-segment the heart followed by a bounding box regression network provided the best compromise between accuracy and simplicity.
- For the first time, a complete benchmark of attention-based segmentation networks involving five different architectures was conducted to evaluate the performance of these methods for the segmentation of left ventricular structures in 2D echocardiography. To this end, several methods proposed in the literature were adapted to our problem.
- A novel segmentation architecture called LU-Net was proposed based on the joint optimization of the localization and the segmentation tasks.
- Our method *i)* outperforms U-Net1, the current best performing deep learning solution on the CAMUS dataset; *ii)* produces among the best results from the tested attention-based networks; *iii)* produces overall segmentation scores lower than the intra observer variability for the epicardial border with 11% of outliers; *iv)* closely reproduces the expert analysis for the end-diastolic and end-systolic left ventricular volumes, with a mean correlation of 0.96; *v)* improves the estimation of the ejection fraction of the left ventricle, with scores that remain slightly higher than the intra observer's ones.

Though the intra-variability remains to be reached for a set of metrics, this study established attention mechanisms as a lead for more robust 2D echocardiographic image segmentation.

Acknowledgment

We would like to thank Dr. Ozan Oktay for his help in the implementation of AG-U-Net. This work was performed within the framework of the LABEX PRIMES (ANR- 11-LABX-0063) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR). The Centre for Innovative Ultrasound Solutions (CIUS) is funded by the Norwegian Research Council (project code 237887).

References

[1] 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 on Echocardiographic Three-Dimensional Ultrasound Segmentation (CETUS), Boston, MIDAS Journal*, 2014, pp. 17–24.

[2] C. Wang and O. Smedby, "Model-based left ventricle segmentation in 3d ultrasound using phase image," in *Proc. MICCAI Challenge on Echocardiographic Three-Dimensional Ultrasound Segmentation (CETUS), Boston, MIDAS Journal*, 2014, pp. 81–88.

[3] 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 on Echocardiographic Three-Dimensional Ultrasound Segmentation (CETUS), Boston, MIDAS Journal*, 2014, pp. 65–72.

[4] 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 on Echocardiographic Three-Dimensional Ultrasound Segmentation (CETUS), Boston, MIDAS Journal*, 2014, pp. 25–32.

[5] 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 on Echocardiographic Three-Dimensional Ultrasound Segmentation (CETUS), Boston, MIDAS Journal*, 2014, pp. 73–80.

[6] 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 on Echocardiographic Three-Dimensional Ultrasound Segmentation (CETUS), Boston, MIDAS Journal*, 2014, pp. 57–64.

[7] 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 on Echocardiographic Three-Dimensional Ultrasound Segmentation (CETUS), Boston, MIDAS Journal*, 2014, pp. 49–56.

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

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

[10] E. Evain, K. Faraz, T. Grenier, D. Garcia, M. De Craene, and O. Bernard, "A pilot study on convolutional neural networks for motion estimation from ultrasound images," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 2020.

[11] E. Smistad, A. Østvik, I. M. Salte, D. Melichova, T. M. Nguyen, K. Haugaa, H. Brunvand, T. Edvardsen, S. Leclerc, O. Bernard, B. Grenne, and L. Løvstakken, "Real-time automatic ejection fraction and foreshortening detection using deep learning," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 2020.

[12] 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 Transactions on Image Processing*, vol. 21, no. 3, pp. 968–982, 2012.

[13] H. Chen, Y. Zheng, J. H. Park, P. Heng, and S. K. Zhou, "Iterative multi-domain regularized deep learning for anatomical structure detection and segmentation from ultrasound images," in *MICCAI*, vol. 9901, 2016, pp. 487–495.

[14] S. Dong, G. Luo, K. Wang, S. Cao, Q. Li, and H. Zhang, "A Combined Fully Convolutional Networks and Deformable Model for Automatic Left Ventricle Segmentation Based on 3D Echocardiography," *BioMed Research International*, vol. 2018, pp. 1–16, 2018.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[16] E. Smistad, A. Østvik, B. O. Haugen, and L. Lovstakken, "2D left ventricle segmentation using deep learning," in *2017 IEEE International Ultrasonics Symposium (IUS)*, 2017, pp. 1–4.

[17] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. de Marvao, T. Dawes, D. P. O'Regan, B. Kainz, B. Glocker, and D. Rueckert, "Anatomically Constrained Neural Networks (ACNNs): Application to Cardiac Image Enhancement and Segmentation," *IEEE Transactions on Medical Imaging*, vol. 37, no. 2, pp. 384–395, 2018.

[18] M. Li, S. Dong, Z. Gao, C. Feng, H. Xiong, W. Zheng, D. Ghista, H. Zhang, and V. H. C. de Albuquerque, "Unified model for interpreting multi-view echocardiographic sequences without temporal information," *Applied Soft Computing*, vol. 88, p. 106049, 2020.

[19] M. Li, C. Wang, H. Zhang, and G. Yang, "MV-RAN: Multiview recurrent aggregation network for echocardiographic sequences segmentation and full cardiac cycle analysis," *Computers in Biology and Medicine*, vol. 120, p. 103728, 2020.

[20] 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.

[21] W. Xue, A. Islam, M. Bhaduri, and S. Li, "Direct multitype cardiac indices estimation via joint representation and regression learning," *IEEE Transactions on Medical Imaging*, vol. 36, no. 10, pp. 2057–2067, 2017.

[22] R. Ge, G. Yang, Y. Chen, L. Luo, C. Feng, H. Zhang, and S. Li, "PV-LVNet: Direct left ventricle multitype indices estimation from 2D echocardiograms of paired apical views with deep neural networks," *Medical Image Analysis*, vol. 58, p. 101554, 2019.

[23] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6450–6458.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 91–99.

[25] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[26] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[27] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, pp. 197 – 207, 2019.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 5998–6008.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[30] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, 2015, pp. 2048–2057.

[31] D. M. Vigneault, W. Xie, C. Y. Ho, D. A. Bluemke, and J. A. Noble, "Omega-net: Fully automatic, multi-view cardiac mr detection, orientation, and segmentation with deep neural networks," *Medical Image Analysis*, vol. 48, pp. 95 – 106, 2018.

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[33] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P. Jodoin, T. Grenier, C. Lartizien, J. D'hooge, L. Lovstakken, and O. Bernard, "RU-Net: A refining segmentation network for 2D echocardiography," in *IEEE International Ultrasonics Symposium (IUS)*, 2019.

[34] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning Where to Look for the Pancreas ," in *Medical Imaging with Deep Learning (MIDL'18)*, 2018.

[35] E. Pesce, S. J. Withey, P.-P. Ypsilantis, R. Bakewell, V. Goh, and G. Montana, "Learning to detect chest radiographs containing pulmonary lesions using visual attention networks," *Medical Image Analysis*, vol. 53, pp. 26 – 38, 2019.

[36] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P. Jodoin, T. Grenier, C. Lartizien, J. D'hooge, L. Lovstakken, and O. Bernard, "Deep learning for segmentation using an open large-scale dataset in 2d echocardiography," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2198–2210, Sep. 2019.

[37] Z. Zhou, M. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *in proc. of Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2018, pp. 3–11.

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

[39] E. D. Folland, A. F. Parisi, P. F. Moynihan, D. R. Jones, C. L. Feldman, and D. E. Tow, "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.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012, pp. 1097–1105.
[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
[43] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2017, pp. 240–248.