IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier xxxx.DOI

ResDUnet: A Deep Learning based Left Ventricle Segmentation Method for Echocardiography

ALYAA AMER[1,2] XUJIONG YE[1], AND FARAZ JANAN.[3]
[1]Computer Science Department, University of Lincoln, Lincoln LN6 7TS, U.K.
[2]Computer Science Department, Arab Academy for Science Technology and Maritime Transport, Cairo 1029, Egypt.
[3]Department of Computing, Imperial college London, London SW7 2BX U.K.

Corresponding author: Alyaa Amer (e-mail: aamer@ lincoln.ac.uk).

ABSTRACT Segmentation of echocardiographic images is an essential step for assessing the cardiac functionality and providing indicative clinical measures, and all further heart analysis relies on the accuracy of this process. However, the fuzzy nature of echocardiographic images degraded by distortion and speckle noise poses some challenges on the manual segmentation task. In this paper, we propose a fully automated left ventricle segmentation method that can overcome those challenges. Our method performs accurate delineation for the ventricle boundaries despite the ill-defined borders and shape variability of the left ventricle. The well-known deep learning segmentation model, known as the U-net, has addressed some of these challenges with outstanding performance. However, it still ignores the contribution of all semantic information through the segmentation process. Here we propose a novel deep learning segmentation method based on U-net, named ResDUnet. It incorporates feature extraction at different scales through the integration of cascaded dilated convolution. To ease the training process, residual blocks are deployed instead of the basic U-net blocks. Each residual block is enriched with a squeeze and excitation unit for channel-wise attention and adaptive feature re-calibration. The performance of the method is evaluated on a dataset of 2000 images acquired from 500 patients with large variability in quality and patient pathology. ResDUnet outperforms state-of-the-art methods with a Dice similarity increase of 8.4% and 1.2% compared to deeplabv3 and U-net, respectively. Furthermore, to demonstrate the impact of each proposed sub-module, several experiments have been carried out with different designs and variations of the integrated sub-modules. We also describe and discuss all technical elements of a deep-learning model via a step-by-step explanation of parameters and methods, while using our left ventricle segmentation as a case study, to explain the application of AI to echocardiographic imaging.

INDEX TERMS Convolutional neural network, Echocardiography, Left ventricle, Segmentation, Convolutional neural network.

I. INTRODUCTION

CARDIOVASCULAR diseases (CVDs) are the number one cause of death worldwide, causing the death of millions of people each year [1]. CVD scans are performed using magnetic resonance imaging (MRI), computed tomography (CT) and echocardiography. Among these, echocardiography is the most common medical imaging modality due to its simplicity, non-invasive nature, and the absence of ionizing radiation [2]. Also, echocardiography provides real-time acquisition with a very high frame rate which is most suitable for assessing the cardiac movement, and in several studies echocardiography has shown comparable results to MRIs [2], [3].

During a clinical exam, a physician ideally performs segmentation of the left ventricle (LV) to examine the cardiac anatomy and provide several clinical measures which illustrate the proper functionality of the heart such as the end-diastolic (ED) and end-systolic (ES) volumes, ejection fraction, LV mass, etc. [3]. Those measures are manually obtained by delineating the LV borders at ED and ES of each cardiac cycle. Despite the advantages of echocardiography and physicians' preference, there are some limitations that make the manual segmentation a very tedious task. Manual delineation is time-consuming and prone to intra and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3122256, IEEE Access

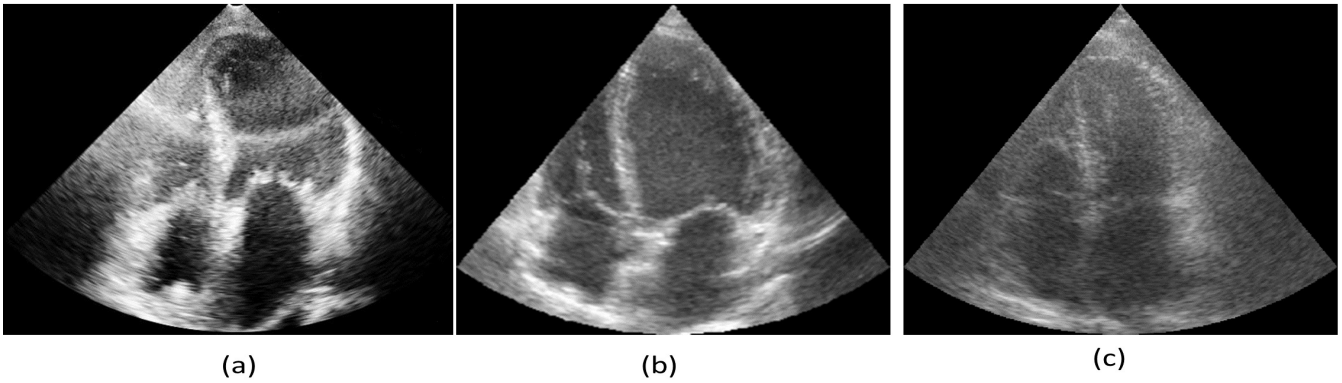

FIGURE 1. Apical four chamber views for echocardiographic images at three different quality levels. (a) Good quality. (b) Fair quality. (c) Poor quality.

inter-observer variability. Those limitations come from the challenging nature of echocardiographic images with low signal-to-noise ratio, motion artifacts, and lower spatial and temporal resolution [4], thus resulting in images with ill-defined LV ventricular wall. Such limitations have motivated the development of an automated segmentation method to help physicians analyze the cardiac anatomy, reduce the clinical examination time and the inter and intra-observer variability. Figure 1 shows an example of the different quality echocardiographic scans.

In this paper, we propose a fully automated LV segmentation method based on a novel deep learning model specially designed and optimized for echocardiographic images. CAMUS dataset [5], a publicly available dataset, is used for training and testing the model. The model leverages its strengths from the well known U-net architecture, along with cascaded dilated convolution, and residual blocks enriched with squeeze and excitation units. The main innovation of our model is:

- The use of residual blocks as the basic building units to extract coarse to fine features from source image. This is done through identity shortcut connections that propagates information from low-level to high-level feature space without the loss of significant details. Moreover, squeeze and excitation units are added to residual blocks for channel-wise attention, adaptive feature recalibration and increasing feature representation power.
- The extraction of features at different scales, unlike the original U-net. This is done by integrating a cascaded dilated convolution module, which increases the receptive field to capture multi-scale information.
- Validating our model using a large dataset of 500 patients with a variability in image quality and patient pathology. We demonstrate its robustness by comparing the results to deeplabv3 and U-net segmentation models.

Worth to mention, we published this automated left ventricle segmentation method, named ResDUnet in [6]. However, here we provide a detailed description for the integrated modules, a clarification for the influence of each module on the whole method, and a step-by-step explanation of the method performance by tuning it's parameters.

The remainder of the paper is structured as follows: Section 2 provides an overview of the related LV segmentation methods. Section 3 gives a detailed description of the methodology of our deep learning model. A brief description of the dataset followed by implementation details and evaluation measures are explained in Section 4. Section 5 details the experimental results and analysis. Finally, we conclude the paper in Section 6.

A. RELATED WORK

Recent advances in the use of echocardiography, not only expanded the scope of application and diagnosis in the conventional sense, but also opened a new era in semi-automated therapy and image-guided interventions. Accordingly, the AI-powered automatic echo image segmentation and it's understanding is an important research question [7]–[9].

Typically, shape priors are used for the LV segmentation. For instance, Chen et al. [10] used active contour constrained with a shape map; however, the method was limited to an amount of a dataset which can not deal with the large variability in the LV shape and size. A fully automatic method using B-spline active surface was proposed by Barbosa et al. [11], which attained the best score in the challenge of cardiac endocardial ultrasound segmentation (CETUS), providing a Dice similarity measure of 0.937. CETUS is a small dataset that consists of 45 Three-dimensional Ultrasound sequences (15 for training, 30 for testing). The B-spline method was later modified by Pedrosa et al. [12] using the same dataset. Their work integrated a shape prior method based on principal component analysis and obtained an average Dice similarity of 0.909 and an average Hausdorff distance of 6.3.

Moreover, Bernard et al. [13], compared the performance of various segmentation methods on the same dataset (CETUS) such as, image-based techniques relying on the assumption of intensity features, motion-based techniques, deformable models, graph cuts, and the B spline explicit active surface technique proposed by Barbosa et al. [11]. Despite the initial success of these methods, they are still insufficient to faithfully delineate boundaries in large datasets with low quality and variability in the LV shapes and size.

Following the recent success of deep learning models to learn multi-scale features for object segmentation, Smistad et al. [14] used U-net architecture with Kalman filter on a small training data of 2D echocardiographic images. It surpassed the performance of deformable models. Furthermore, Oktay et al. [15] used convolution neural network (CNN) to propose an approach named anatomically constrained neural network (ACNN), that uses auto-encoder to fit non-linear compact representation of the LV structure. They evaluated their approach on the CETUS dataset, achieving a Dice similarity measure of 0.912 and an average Hausdorff distance of 7.0.

Leclerc et al. [5] introduced a large public dataset, CAMUS (Cardiac Acquisition for Multi-structure Ultrasound Segmentation) and benchmarked the performance of two different implementations of U-net architecture with different hyper-parameter choices. One optimized for speed and the other optimized for accuracy. Their performance was evaluated against the approaches suggested by Oktay et al. [15], Smistad et al. [14], and Barbosa et al. [11]. Leclerc et al. [5], showed that the aforementioned U-net implementations outperformed the sophisticated deep learning approach proposed by Oktay et al. [15], and the B-spline prior approach suggested by Barbosa et al. [11]. Leclerc's U-net approach was able to achieve a Dice similarity measure of 0.939, achieving state-of-the-art performance on CAMUS dataset which consists of 500 patients, each with two different views, apical two-chamber and apical four-chamber view. Also, Azarmehr et al. [7] compared the performance of the two U-net implementations proposed by Leclerc et al. [5] against the original U-net model [16]. They trained their model on a dataset of 992 images acquired from 61 patients, they concluded that the original U-net provides better performance on their dataset in comparison with the modified U-net implementations provided by Leclerc et al. [5].

To the best of our knowledge, ResDUnet is the first method to combine the strengths of different modules, i.e. dilated convolution, residual blocks, and squeeze and excitation unit. Those modules are all integrated into one framework to help capture and aggregate the LV's multi-scale features for more accurate and robust segmentation across different quality images with varying LV sizes.

II. METHODOLOGY

In this section we will provide a step-by-step explanation of our method, ResDUnet, depicted in Figure 2. ResDUnet is built on the U-net architecture with significant modifications to adapt to the nature of echocardiographic images and their applications. In principle, we have replaced the basic U-net building blocks with alternative residual blocks enriched with squeeze and excitation units, along both the encoding and decoding path. Along the encoding path, We used residual blocks integrated with squeeze and excitation units and extracted multi-scale features using cascaded dilated convolution with varying dilation rates. Along the decoding path, feature maps are upsampled along with regular convolution for precisely locating information, where the image size gradually increases, and the depth gradually decreases. Implementation details of each step are explained as follows.

A. RESIDUAL BLOCKS

Inspired by the success of U-net [16] and deep residual learning [17], we combine the strengths of both into a single model. It is known that deeper networks provide better performance [17], nevertheless, training a deep U-net architecture is difficult, especially for a limited amount of training data. To some degree, we can overcome data shortage by deploying a pre-trained network with fine-tuning [18]. He et al. [17] introduced residual blocks that share the same idea of concatenating the input (identity short-cut) and propagating the low fine details. This enhances network performance without the need to go deeper or deploy a pre-trained network. Accordingly, we replace the basic U-net units (Figure 3(a)) with residual blocks (Figure 3(c)). Those residual blocks further contribute in feature propagation at the encoding and decoding paths. Also, fine feature details are propagated to coarser layers through convolution with stride instead of pooling (which is originally used in U-net). Each residual unit is defined as:

$$y_i = h(x_i) + F(x_i, W_i) \tag{1}$$

$$x_{i+1} = f(y_i) \tag{2}$$

where x_i and x_{i+1} are the input and output of the *i-th* residual unit, $F(.)$ is the residual function, $f(y_i)$ is the activation function and $h(x_i)$ is an identity mapping function. ResDUnet uses the pre-activation design illustrated in figure 3(c), which is the most optimized combination as investigated in [19].

B. SQUEEZE AND EXCITATION UNIT

Within each residual block, squeeze and excitation (SE) unit is added before performing identity mapping. Our model is mechanized this way since integrating SE blocks into convolution neural networks helps boost the end-to-end network performance, especially when added to residual and inception networks [17]. The SE unit allows the network at earlier layers to excite informative features and strengthen the shared low-level feature representation. Moreover, at deeper layers, it responds differently to inputs in a highly class-specific manner. Please refer to [20] for more details and experimental derivation of the SE units, nevertheless, for the sake of completeness we will explain it briefly.

The structure of the SE unit is shown in Figure 3(b). Each unit performs two fundamental operations on the input feature map; squeeze for global information embedding and excitation for feature recalibration. First, features are passed through a squeeze operation which produces a channel descriptor by aggregating feature maps across their spatial dimensions H×W. This is done by global average pooling

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3122256, IEEE Access

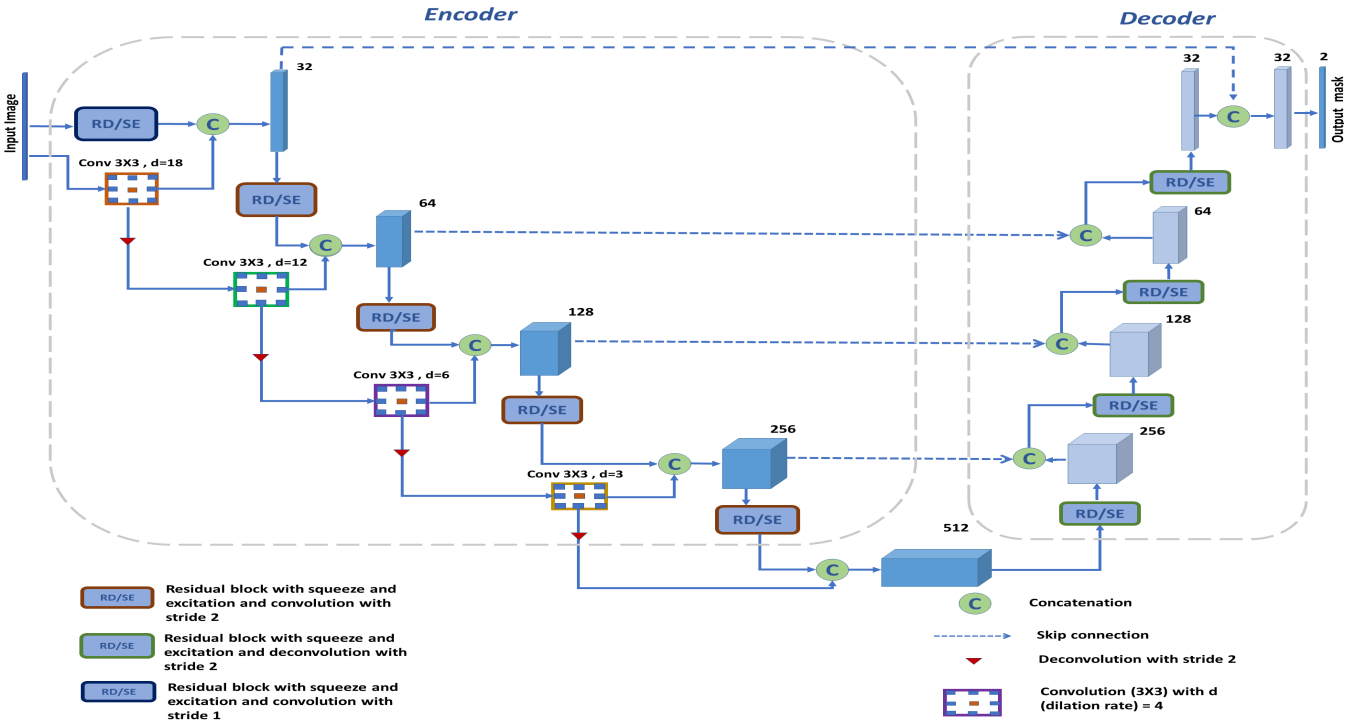

FIGURE 2. The architecture of the proposed ResDUnet model specifically designed and optimized for echocardiographic images.

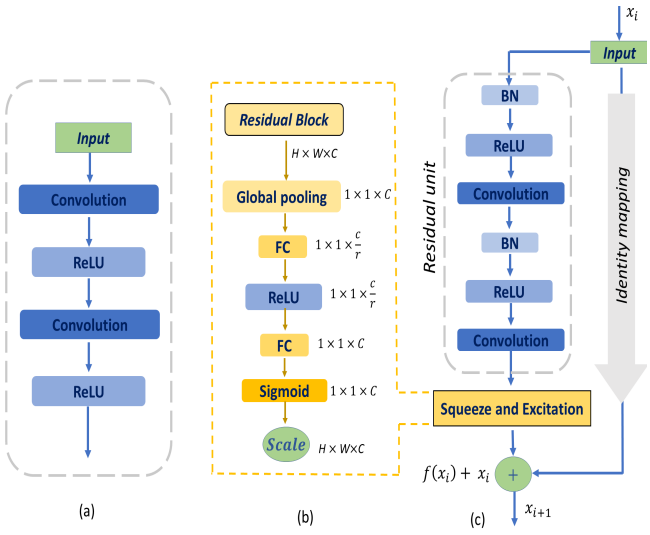

FIGURE 3. Building blocks. (a) U-net basic building block. (b)Squeeze and excitation unit. (c) Residual block with squeeze and excitation added before identity mapping (Standard-SE).

to generate channel-wise statistics that embeds global distribution of channel-wise feature responses (C). This allows information from the network's global receptive field to be used by all its layers. Second, the aggregated channels are passed to an excitation operation, a simple self-gating mechanism using sigmoid activation function. Excitation is added to fully capture channel-wise dependencies and learn non-linear, non mutual and non-exclusive relationships [20].

After global average pooling, there is a bottleneck with two fully connected (FC) layers with a reduction ratio (r). The output of the whole SE unit is obtained by rescaling the transformed feature maps with the activations, which act as weights adapted to the input features [20]. Roy et al. [21] have extensively studied the use of SE units in three different fully convolutional architectures (U-Net, SD-Net and FC-DenseNet) on three different medical image applications. They concluded that each SE unit has contributed in increasing the segmentation performance by boosting meaningful features and suppressing weak ones. Also, they illustrated that the substantial increase in segmentation accuracy comes with a negligible increase in model complexity, concluding that squeeze and excitation can be a crucial component for fully convolutional in many medical applications.

C. CASCADED DILATION

As discussed previously, U-net provides good performance through skip connections. However, this direct concatenation process ignores the contribution of all semantic strengths in the segmentation process. It does not take into consideration various scales of the feature map. Moreover, in the U-net architecture, the feature resolution is gradually lost in the process of sub-sampling layers throughout the network. This results in coarse features that miss the details of very small objects that are difficult to recover even with the effort of skip connections [16], [18]. It is a common knowledge that in a typical cardiac cycle, the LV changes in size and shape as a result of contraction and relaxation. Therefore, to deal with the problem of LV shape variability, we adopted cascaded

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3122256, IEEE Access

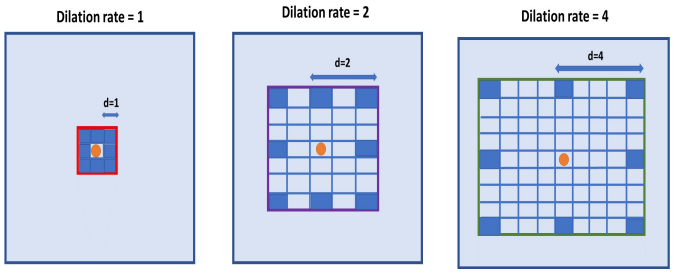

FIGURE 4. Dilated convolution with kernel size 3×3 and different dilation rates. Dilated convolution with d = 1, corresponds to normal convolution.

dilated convolution with variable dilation rates. This enables the extraction of a multitude of features compensating for variability in the appearance of LV [22], [23].

In comparison with conventional convolution, dilated convolution is able to achieve a larger receptive field size, where 3×3 kernel with a dilation rate of 2 results in a receptive field size equal to that of a 5×5 kernel, without increasing the network complexity [24]. Dilated convolution expands kernel weight alignments with a dilation rate (d); increasing this factor provides more sparse weights. The size of the kernel and the receptive field increases accordingly and enlarges the field of filter to incorporate global multi-scale context. For a two-dimensional feature map x, a filter w, and for each location i on the output y, dilated convolution is explained as in [24]:

$$y[i] = \sum_{k} x[i + d.k]w[k] \tag{3}$$

we assume that the dilation rate d is equivalent to the stride that we use to sample the input signal. This process resembles convolving the input x with the up-sampled filters produced by inserting (d-1) zeros between two consecutive filter values alongside each spatial dimension. Thereby, using dilated convolution rates, we can adjust the filter's field of view. Figure 4 shows how the receptive field and kernel size increase by increasing the dilation factor, which helps incorporate the global context of the image.

Although dilated convolution provides a large receptive field size for more global context, the segmentation mask generated by simply one dilation rate throughout the whole segmentation process still does not cover all semantic strengths. However, stacking dilated convolution layers with different dilation rates will further increase the receptive field and boost the segmentation performance [23], [24]. This dilation strategy has been utilized in medical image domain as seen in [25], where the authors developed an approach for segmenting multiple organs from axial CT images. They employed several dilated convolution with different dilation rates in the bottleneck layer (the lowest layer of the encoder-branch) of the U-net model. They claim that performance gains were achieved when they used dilated convolution to capture both local and global information.

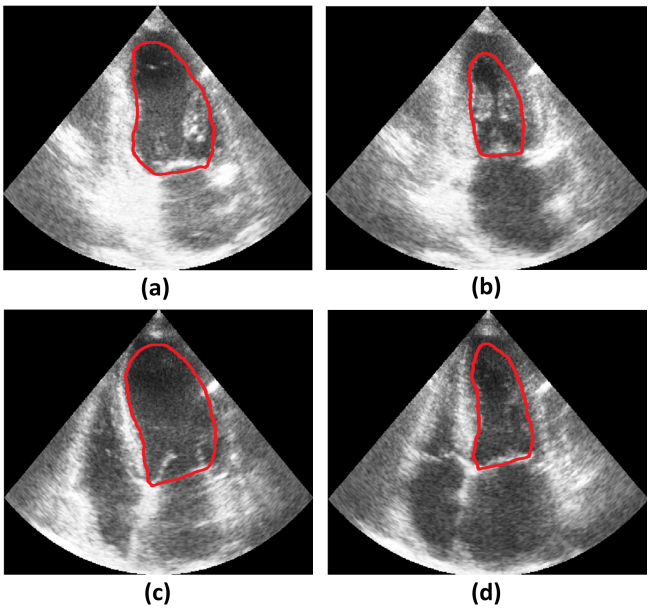

FIGURE 5. Expert annotation of the LV from two different views at two phases. Apical two-chamber (A2C) at (a) end-diastole phase and (b) end-systole phase. Apical four-chamber (A4C) at (c) end-diastole phase and (d) end-systole phase.

In short, ResDUnet has a cascaded dilated convolution topology with varying rates along the encoding path, inspired from [24], to cover all scales in the image and further propagate this information along skip connections to the corresponding high level semantics.

III. MATERIALS AND EVALUATION METRICS

In this section, we first provide a detailed description for the dataset used to evaluate the performance of the proposed model. Then we discuss the implementation details followed by description of evaluation measures used in our study.

A. DATASET

We used a publicly available dataset called CAMUS (Cardiac Acquisition for Multi-structure Ultrasound Segmentation) [5], which contains 2D echocardiographic images for 500 patients, annotated by two qualified clinicians. The full dataset is acquired at the University Hospital of St Etienne (Fran), from GE Vivid E95 ultrasound scanners (GE Vingmed Ultrasound, Horten Norway), with a GE M5S probe (GE Healthcare, US). For each patient, 2D apical four-chamber and two-chamber view sequences were exported from EchoPAC analysis software (GE Vingmed Ultrasound, Horten, Norway). Figure 5 shows an example of the two different frames from echocardiographic images annotated by the expert at both, the end-diastole and end-systole phases. These are the frames that are typically used for measuring clinical indices. The size of the images varies from one patient to another, with an average resolution of 500×600. The dataset involves a wide variability of acquisition settings to imitate the real clinical situation by including all types of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3122256, IEEE Access

images regardless of their quality, i.e., some images had ill-defined, poorly-visible or vanished ventricular wall, which requires further expert analysis. The dataset also includes images for patients with various pathological conditions, i.e. pulmonary thromboembolism, pulmonary, hypertension, cardiomyopathies, etc.

To add robustness and reduce over-fitting, we augmented the dataset using the image augmentation library 'Augmenter' [26]. We applied vertical transformation to 30% of the dataset, horizontal transformation to 30%, and the rest were vertically and horizontally transformed. In total, there are 4000 images, 3000 images for training with 20% for validation, while the remaining 1000 images for testing.

B. IMPLEMENTATION DETAILS

We implemented our model in Python environment with keras 2.2.4 and Tensorflow 1.15. We used the Adam optimizer at 1000 epochs, with a learning rate initially set to 10^{-3} and reduced by a factor of 10 every 200 epochs. Dice loss is used as the network's objective function. The testing process for 1000 images took 63.2 seconds, while the training process took 23 hours on a computer using a mid-range NVIDIA GeForce with RTX 2070 GPU.

For our results to be comparable to Leclerc et al. [5] on the CAMUS dataset, we followed their approach and downsized the images to a resolution of 256×256. Our implementation uses convolution layers with a kernel size of 3×3. We also apply batch normalization and Rectified Linear Unit (ReLU) as an activation function, and the max-pooling layers throughout the encoding path are replaced by deconvolution with a stride of 2. The network depth is 5 layers. We used different dilation rates across the cascaded dilated convolution in the encoding path. We started the first dilated convolution with a rate of 18 and reduced it till 3 at the bottom layer, where the spatial information is most compressed.

To provide effective convolution, we followed the same tuning of dilation rates in [27] in an inverted way to map the varying feature map size as we go down the encoding path. Following the last batch normalization within each residual block, we added a squeeze and excitation unit. Each squeeze and excitation unit has a reduction ratio of 8, which provides the lowest overall error with Resnet [20].

C. EVALUATION MEASURES

In this section we illustrate the statistical measures used to compare our predicted LV region (S_P) against the ground truth manually drawn by the experts (S_M). We used three segmentation performance metrics: Dice similarity coefficient (DSC) [28], Hausdorff distance [29], and mean absolute distance (MAD) [30]. The DSC measure is described as the overlap between S_P region and S_M region, and is defined as:

$$DSC = \frac{2(S_P \cap S_M)}{S_P + S_M} \tag{4}$$

The higher the value for DSC, the more it matches to the ground truth. When the DSC reaches a value of 1, this indicates a complete overlap. The HD estimates how far the two contours are from each other. Precisely, it is the largest of all distances from a point in one set to the closest point in the other set. If A and B represent predicted and manual contours, respectively, HD from A to B is the maximum function defined as:

$$\begin{aligned} H(A,B) &= max(h(A,B), h(B,A)), \\ h(A,B) &= max_{a\in A}(min_{b\in B}(d(a,b))), \\ h(B,A) &= max_{b\in B}(min_{a\in A}(d(b,a))) \end{aligned} \tag{5}$$

A lower value of the HD represent a higher match between the two contours, whereas d(:) is the euclidean distance between any two points. MAD measure is the average of all distances between the points along two contours, and is defined as:

$$MAD = \frac{1}{N_P}\sum_{i=1}^{N_P} |d(a,b)| \tag{6}$$

N_P represents the number of points in the predicted contour. A and B represent predicted and manual contours.

IV. EVALUATION RESULTS AND DISCUSSION

In this section, we compared the performance of ResDUnet with state-of-the-art approaches. Specifically, a comparison is made with deeplabv3 [24]. This model employed cascaded dilated convolution with upsampled filters and doubling the dilation rates to encode multi-scale information. Also, a comparison is made with the model implemented in [5]. This model is a basic U-net model optimized for accuracy by adding batch normalization layers and going deeper, starting with 32 feature maps and ending with 512 feature maps at the bottom layer. When this basic U-net model was applied on CAMUS dataset in [5], it has outperformed U-net++ by 1.2%, Stacked Hourglasses method (SHG) by 0.5% and ACNN by 0.7%. Also, it has offered the best compromise between the network size and performance for the task of 2D echocardiographic image segmentation, requiring less parameters than SHG and U-Net++ methods and less training time than ACNN.

We also investigated the effect of adding each module to the basic U-net architecture. Then, we studied the impact of parameters tuning in the cascaded dilated convolution and squeeze and excitation unit. Finally, we studied the model performance when the SE unit is integrated at different locations with respect to the residual unit.

A. COMPARISON WITH DEEPLABV3 AND U-NET

The segmentation performance of the proposed model in comparison with U-net and deeplabv3 is illustrated in Table 1. It can be seen that ResDUnet outperforms deeplabv3 and U-net by 8.4% and 1.2%, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3122256, IEEE Access

A higher DSC measure is seen in U-net in comparison with deeplabv3. This is due to the concatenation of low-level features with the corresponding high-level features through skip connections. ResDUnet outperformed the results achieved by U-net due to the integration of multi-scale features along the skip connections and the encoding path. Also, the presence of the squeeze and excitation unit integrated within each residual block has enriched the feature extraction process with more informative features and minimized less informative ones.

Method	DSC	HD	MAD
U-net	0.939 ± 0.043	5.3 ± 3.6	1.6 ± 1.3
Deeplabv3	0.867 ± 0.113	6.5 ± 4.9	1.8 ± 2.1
ResDUnet	$\mathbf{0.951 \pm 0.030}$	$\mathbf{4.5 \pm 1.2}$	$\mathbf{1.4 \pm 1.2}$

TABLE 1. Statistical evaluation for ResDUnet in comparison with other segmentation models (mean value $\pm$ standard deviation). HD and MAD are measured in mm.

Figure 6 shows a qualitative comparison between the predicted contours from ResDUnet versus U-net and deeplabv3. The manual segmentation by the experts is depicted using the contour drawn in 'green' color and the predicted contour is depicted in 'red'. The predicted contours and the experts contour are both extracted from their respective binary masks using a boundary visualization function in matlab, named visboundaries(.). Here, to compare the performance of the three models, we selected three different quality images, which we visually considered as 'good', 'fair' and 'poor' in terms of how well defined the LV wall is.

In the good case (Figure 6(a)), where the LV wall is well defined, the three models provided comparable segmentation in reference to manual delineation. Nevertheless, ResDUnet provides overall the best performance. In the case of 'fair' quality image (Figure 6(b)), the two models (Deeplab and U-net) failed to precisely outline the LV in comparison with ResDUnet, which outperformed. Finally, in the 'poor' quality image (Figure 6(c)), where the ventricular wall is very ill-defined, deeplabv3 provided a very poor segmentation, whereas ResDUnet offers comparable results to U-net. However, a better alignment with experts annotation is seen in ResDUnet.

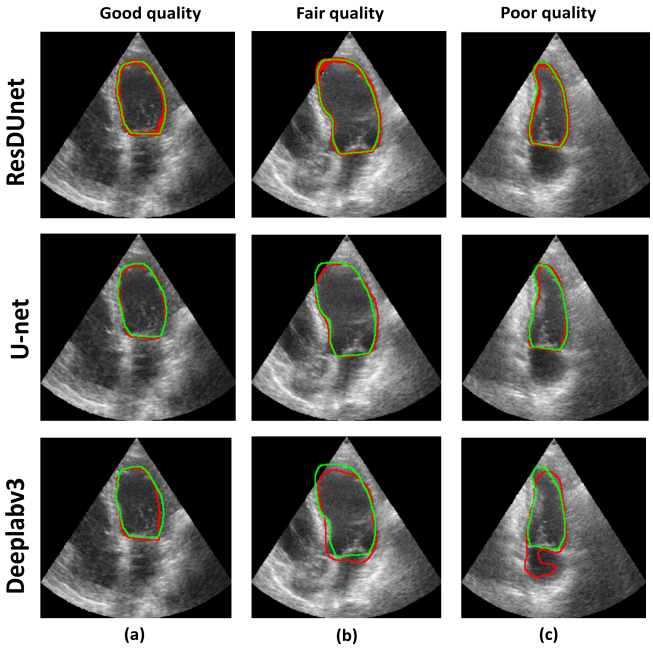

FIGURE 6. LV segmentation by ResDUnet, U-net ,and Deeplabv3. Red contours present predicted segmentation. Green contours present gold standard. (a) Apical four-chamber at end-systole phase. (b) Apical four-chamber at end-diastole phase. (c) Apical two-chamber at end-diastole phase. Image quality is shown above each column.

B. EVALUATION OF MODULES ADDITION

In this section, we studied the contribution of each module to ResDUnet. First, we showed the effect of adding residual blocks to the original U-net model. Second, we illustrated the effect of enriching each residual unit with squeeze and excitation unit. Finally, we showed the model performance after adding the cascaded dilation module. From the segmentation results shown in Table 2, it is observed that adding residual units has increased the model performance. This enhanced performance is due to the presence of identity short-cut mapping in each residual block which allows the flow of information from initial layers to last layers and reduces the problem of vanishing and exploding gradients. The model performance was further enhanced when each residual block was integrated with SE unit. The SE unit has contributed in increasing the feature representation power through excitation operation and suppressing less important information through squeeze operation. The model performance was further boosted when cascaded dilation was added along the encoding path. This is due to the large and variable receptive field integrated in the cascaded scheme to capture the multi-scale features.

Method	DSC	HD	MAD
RD	0.941 ± 0.139	5.4 ± 0.9	3.1 ± 1.7
RD/SE	0.942 ± 0.143	5.3 ± 1.9	2.2 ± 1.8
RD/SE/CD	$\mathbf{0.951 \pm 0.030}$	$\mathbf{4.5 \pm 1.2}$	$\mathbf{1.4 \pm 1.2}$

TABLE 2. Segmentation results after adding each module; RD: residual units; SE: squeeze and excitation; CD: cascaded dilation). HD and MAD are measured in mm.

C. EVALUATION OF SQUEEZE AND EXCITATION UNIT

In this section we investigated the model performance by changing the reduction ratio (r), mentioned in section II-B, and we examined the effect of changing the location of the SE unit in reference to the residual unit.

This reduction ratio (r) changes the capacity and computational cost of each block in the network. In table 3 we represent the Dice similarity coefficient for different values of the reduction ratio at r = 2, 4, 8, 16 and 32. We notice that the model segmentation performance is robust for various

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3122256, IEEE Access

Ratio	DSC	Parameters
2	0.943	$46.3M$
4	0.942	$45.9M$
8	0.951	$45.8M$
16	0.933	$45.7M$
32	0.949	$45.6M$

TABLE 3. Segmentation results and parameter sizes for different ranges of (r) in each SE unit.

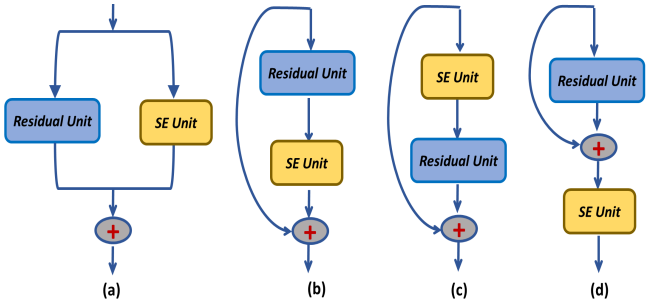

FIGURE 7. Integration of squeeze and excitation unit with residual unit. (a) SE-Identity. (b) SE-Standard. (c) SE-Pre. (d) SE-Post.

values of r, however, a performance drop is seen at r =16. Therefore, we chose a reduction ratio of 8, since it provides a fair balance between the segmentation accuracy and the model complexity, with a parameter size of 45.8 M. This parameter size is comparable to the parameter size seen when using a reduction ratio of 4, 16, and 32, and has increased to 46.3 M for a reduction ratio of 2. It is important to note that users of squeeze and excitations units, may decide to use a different reduction ratios for each base architecture, depending upon the various functionalities of each layer in the networks as confirmed in [20].

Integration	DSC
Post-SE	0.947
Pre-SE	0.930
SE-identity	0.963
Standard-SE	0.951

TABLE 4. Segmentation results for different integrations of the SE unit with the residual unit

We further conducted four experiments that provide an ablation study for integrating the SE unit within each residual block. The four experiments include (1) Post-SE block: SE unit is added after the addition of identity mapping. (2) Pre-SE block: SE unit is added before the residual unit. (3) SE-identity block: where both SE and residual units operate in parallel. (4) Standard-SE block (the integration we adopted in our model): the SE unit is added right after the residual unit and before the addition of the input. These model variations are shown in Figure 7. Their influence on the overall segmentation performance is illustrated in Table 4. It is noticed that the Post-SE and Pre-SE provide comparable results, and slight increase in performance is seen in the adopted integration (Standard-SE), which implies that performance gains achieved by SE units is robust to their location. SE-identity, has provided the best performance out of the three other integrations, but that comes with an increase in the model complexity with more than 10 M trainable parameters compared to the other three designs. Therefore, in ResDUnet we used the Standard-SE integration to achieve the best trade off between segmentation accuracy and model complexity.

D. EVALUATION OF CASCADED DILATION

In this section we investigate the effect of changing the dilation rate on the model segmentation performance using two cascaded dilation schemes inspired from [24] and [31].

As mentioned earlier in section II-C, the dilation rate defines the spaces between kernel weights and help capture a larger global, multi-scale context. Therefore, it helps in resolving the shape variability in the LV captured at both the end-diastole and end-systole phases.

Dilation Rates	DSC
16,8,4,2	0.936
18,12,6,3	0.951

TABLE 5. Segmentation results for different dilation rates integrated in the cascaded dilation module

Table 5 shows the segmentation accuracy for two different cascaded dilation rates of 16/8/4/2 and 18/12/6/3. The largest rate (i.e. 16 or 18) at both experiments is used for the uppermost layer while the smallest rate (i.e. 2 or 3) is used for the lowermost layer. It is noted that using cascaded dilation rates of 18/12/6/3 has increased the segmentation accuracy by 1.5%. This better performance is due to the larger rates that in turn provides a larger receptive field. To explain more, the smallest dilation rate of 3 with a kernel size of 3×3 maps to a kernel size of 7, and a dilation rate of 6 with a kernel size of 3×3 maps to a kernel size of 13. Adding these two dilation rates together in a cascaded fashion will further enlarge the receptive field, providing a kernel size of 19×19. This proves that dense connection between cascaded dilation layers can compose a feature pyramid with much denser scale diversity, which helps overcome variability in the LV size and shape.

V. CONCLUSION

In this paper, we proposed a robust and simple deep learning model, ResDUnet, specifically designed and optimized to segment the LV from echocardiographic images. We have demonstrated that performance gains can be achieved when state-of-the-art U-net is integrated with innovative concepts such as residual blocks, squeeze and excitation units, and dilated convolution. Despite the challenging nature of echocardiographic images with ambiguous ventricular wall and variability in shape and size of the LV. ResDUnet has provided promising results with a Dice similarity measure of 95.1% in-comparison with expert's annotation, and has outperformed the well-known deep learning methods, Deeplabv3 and U-

net by 8.4% and 1.2%, respectively. The model performance was evaluated on a dataset of 2000 images taken from a publicly available dataset (CAMUS). We have conducted several experiments with different designs and variations for the integrated sub-modules. We concluded that the exceptional performance of ResDUnet is achieved by integrating a cascaded dilated module that compensates for occlusions, contraction and relaxation of the LV. Moreover, it enables the increase of kernel size without added complexity. This results in enlarged field of view of the filter, to incorporate greater global and multi-scale context. Furthermore, residual blocks improved the model performance, through the concatenation of the input layer that helps to propagate the low fine details and passing them through the skip connections. Future studies will include increasing the size of the dataset, with different augmentation techniques. We will further investigate the performance of the model in a clinical setting, for instance, to measure the left ventricle ejection fraction, which is highly recommended by physicians to assess the heart functionality. We would also like to investigate the model performance on other different modalities.

REFERENCES

[1] W. H. Organization, "Cardiovascular diseases," 2020.

[2] D. Muraru, L. P. Badano, G. Piccoli, P. Gianfagna, L. Del Mestre, D. Ermacora, and A. Proclemer, "Validation of a novel automated border-detection algorithm for rapid and accurate quantitation of left ventricular volumes based on three-dimensional echocardiography," European journal of echocardiography, vol. 11, no. 4, pp. 359–368, 2010.

[3] H. P. Kühl, Left ventricular and left atrial function, pp. 59–78. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015.

[4] P. S. Douglas, J. M. DeCara, R. B. Devereux, S. Duckworth, J. M. Gardin, W. A. Jaber, A. J. Morehead, J. K. Oh, M. H. Picard, S. D. Solomon, et al., "Echocardiographic imaging in clinical trials: American society of echocardiography standards for echocardiography core laboratories: endorsed by the american college of cardiology foundation," Journal of the American Society of Echocardiography, vol. 22, no. 7, pp. 755–765, 2009.

[5] S. Leclerc, E. Smistad, J. Pedrosa, A. stvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, et al., "Deep learning for segmentation using an open large-scale dataset in 2d echocardiography," IEEE Trans. on medical imaging, 2019.

[6] A. Amer, X. Ye, M. Zolgharni, and F. Janan, "Resdunet: Residual dilated unet for left ventricle segmentation from echocardiographic images," in 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 2019–2022, IEEE, 2020.

[7] N. Azarmehr, X. Ye, F. Janan, J. P. Howard, D. P. Francis, and M. Zolgharni, "Automated segmentation of left ventricle in 2d echocardiography using deep learning," arXiv preprint arXiv:2003.07628, 2020.

[8] A. K. Savaashe and N. V. Dharwadkar, "A review on cardiac image segmentation," in 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 545–550, IEEE, 2019.

[9] Z. Jian, X. Wang, J. Zhang, X. Wang, and Y. Deng, "Diagnosis of left ventricular hypertrophy using convolutional neural network," BMC Medical Informatics and Decision Making, vol. 20, no. 1, pp. 1–12, 2020.

[10] Y. Chen, H. D. Tagare, S. Thiruvenkadam, F. Huang, D. Wilson, K. S. Gopinath, R. W. Briggs, and E. A. Geiser, "Using prior shapes in geometric active contours in a variational framework," International Journal of Computer Vision, vol. 50, no. 3, pp. 315–328, 2002.

[11] 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," MIDAS J, vol. 10, 2014.

[12] J. Pedrosa, S. Queirs, O. Bernard, J. Engvall, T. Edvardsen, E. Nagel, and J. D'hooge, "Fast and fully automatic left ventricular segmentation and tracking in echocardiography using shape-based b-spline explicit active surfaces," IEEE Trans. on medical imaging, vol. 36, no. 11, pp. 2287–2296, 2017.

[13] O. Bernard, J. G. Bosch, B. Heyde, M. Alessandrini, D. Barbosa, S. Camarasu-Pop, F. Cervenansky, S. Valette, O. Mirea, M. Bernier, P.-M. Jodoin, J. S. Domingos, R. V. Stebbing, K. Keraudren, O. Oktay, J. Caballero, W. Shi, D. Rueckert, F. Milletari, S.-A. Ahmadi, E. Smistad, F. Lindseth, M. van Stralen, C. Wang, O. Smedby, E. Donal, M. Monaghan, A. Papachristidis, M. L. Geleijnse, E. Galli, and J. D'hooge, "Standardized evaluation system for left ventricular segmentation algorithms in 3d echocardiography," IEEE transactions on medical imaging, vol. 35, p. 967—977, April 2016.

[14] E. Smistad and F. Lindseth, "Real-time tracking of the left ventricle in 3d ultrasound using kalman filter and mean value coordinates," Medical Image Segmentation for Improved Surgical Navigation, p. 189, 2014.

[15] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'Regan, et al., "Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation," IEEE transactions on medical imaging, vol. 37, pp. 384–395, 2017.

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

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

[18] J. Darrell and U. Berkeley, "Fully convolutional networks for semantic segmentation," UC Berkeley, 2014.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," 2016.

[20] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," 2017.

[21] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks," IEEE transactions on medical imaging, vol. 38, no. 2, pp. 540–549, 2018.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," 2016.

[23] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," 2016.

[24] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, 2017.

[25] S. Vesal, N. Ravikumar, and A. Maier, "A 2d dilated residual u-net for multi-organ segmentation in thoracic ct," arXiv preprint arXiv:1905.07710, 2019.

[26] M. D. Bloice, P. M. Roth, and A. Holzinger, "Biomedical image augmentation using augmentor," Bioinformatics, vol. 35, no. 21, pp. 4522–4524, 2019.

[27] Y. Wang, B. Liang, M. Ding, and J. Li, "Dense semantic labeling with atrous spatial pyramid pooling and decoder for high-resolution remote sensing imagery," Remote Sensing, vol. 11, no. 1, p. 20, 2019.

[28] L. R. Dice, "Measures of the amount of ecologic association between species," Ecology, vol. 26, no. 3, pp. 297–302, 1945.

[29] J. Henrikson, "Completeness and total boundedness of the hausdorff metric," MIT Undergraduate Journal of Mathematics, vol. 1, pp. 69–80, 1999.

[30] S. Kolassa, W. Schütz, et al., "Advantages of the mad/mean ratio over the mape," Foresight: The International Journal of Applied Forecasting, no. 6, pp. 40–43, 2007.

[31] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3684–3692, 2018.