ResDUnet: Residual Dilated UNet for Left Ventricle Segmentation from Echocardiographic Images

Alyaa Amer[1,2], Xujiong Ye[1], Massoud Zolgharni[3], Faraz Janan[1]

***Abstract*— Echocardiography is the modality of choice for the assessment of left ventricle function. Left ventricle is responsible for pumping blood rich in oxygen to all body parts. Segmentation of this chamber from echocardiographic images is a challenging task, due to the ambiguous boundary and inhomogeneous intensity distribution. In this paper we propose a novel deep learning model named ResDUnet. The model is based on U-net incorporated with dilated convolution, where residual blocks are employed instead of the basic U-net units to ease the training process. Each block is enriched with squeeze and excitation unit for channel-wise attention and adaptive feature re-calibration. To tackle the problem of left ventricle shape and size variability, we chose to enrich the process of feature concatenation in U-net by integrating feature maps generated by cascaded dilation. Cascaded dilation broadens the receptive field size in comparison with traditional convolution, which allows the generation of multi-scale information which in turn results in a more robust segmentation. Performance measures were evaluated on a publicly available dataset of 500 patients with large variability in terms of quality and patients pathology. The proposed model shows a dice similarity increase of 8.4% when compared to deeplabv3 and 1.2% when compared to the basic U-net architecture. Experimental results demonstrate the potential use in clinical domain.**

I. INTRODUCTION

Segmentation of the left ventricle (LV) is an inevitable task for examining the cardiac anatomy. According to many studies, echocardiography is the most preferred modality by physicians, due to its non-invasive nature, low cost, and high performance at a real time acquisition [1]. However, echocardiographic images have some limitations such as low signal to noise ratio, low spatial and temporal resolution, and motion artefacts [2], where in some cases, the ventricular wall might disappear, which mislead the delineation process. Different quality images are illustrated in figure1. Those limitations make the manual segmentation task prone to observer variability. Such limitations have motivated the contribution of an automated method for helping physicians in analysing the cardiac anatomy, and reducing the time needed to examine each patient, especially for a large number of frames taken from different views.

The problem of automatic LV segmentation in echocardiographic images has been exploited in the literature [3], [4], [5], [6]. Following the success of deep learning models to learn multi-level features for object segmentation, Oktay et al. [7], used convolution neural network to propose an approach named anatomically constrained neural network (ACNN), which uses auto-encoder to fit non-linear compact representation of the left ventricle structure, achieving a dice similarity measure of 0.912. Moreover, Sarah et al. [8] have introduced a large public dataset and benchmarked the performance of two different implementations of U-net architecture. Those two U-net implementations achieved a dice similarity measure of 0.939, outperforming the sophisticated deep learning approach proposed by Oktay et al. [7], and the B-spline explicit active surface approach proposed by Barbosa et al. [9].

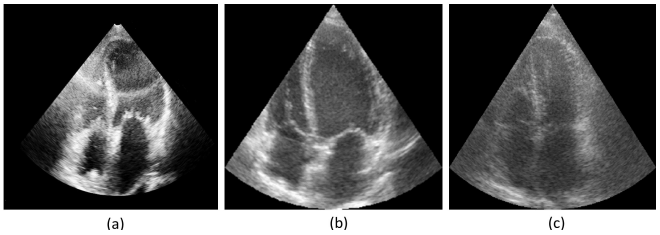

Fig. 1. Echocardiographic images in apical four chamber view at different quality. (a) Good quality. (b) Fair quality. (c) Poor quality.

In this paper, we propose a deep learning model named ResDUnet to automatically segment the LV from echocardiographic images. The model leverage the strengths of U-net, cascaded dilated convolution, and residual blocks enriched with squeeze and excitation. This model brings us the following gains: *i)* The skip connections in U-net and within the residual blocks help the propagation of information from low level features to high level features without loss of many details. *ii)* Using residual block as the building unit, to ease the training process and help extract coarse and fine features from the source image. Also, squeeze and excitation units are added to each residual block for channel-wise attention, adaptive feature recalibration and increasing feature representation power. *iii)* Since U-net ignores extracting features at different scales, we integrated a cascaded module of dilated convolution to provide global and multi-scale feature extraction.

The paper is organized as follows. Section 2 describes the methodology of the proposed model. Section 3 describes the dataset used and implementation details. Evaluation results and discussion are illustrated in section 4. Finally, conclusion is demonstrated in section 5.

[1]Alyaa Amer, Xujiong Ye, Faraz Janan is with School Computer Science, University of Lincoln, Lincoln, UK. aamer@lincoln.ac.uk, xye@lincoln.ac.uk, fjanan@lincoln.ac.uk

[2]Alyaa Amer is with Computer Engineering Department, Arab academy for Science and technology, Alexandria, Egypt.

[3]Massoud Zolgharni is with School of Computing and Engineering, University of West London, London, UK. massoud.zolgharni@uwl.ac.uk

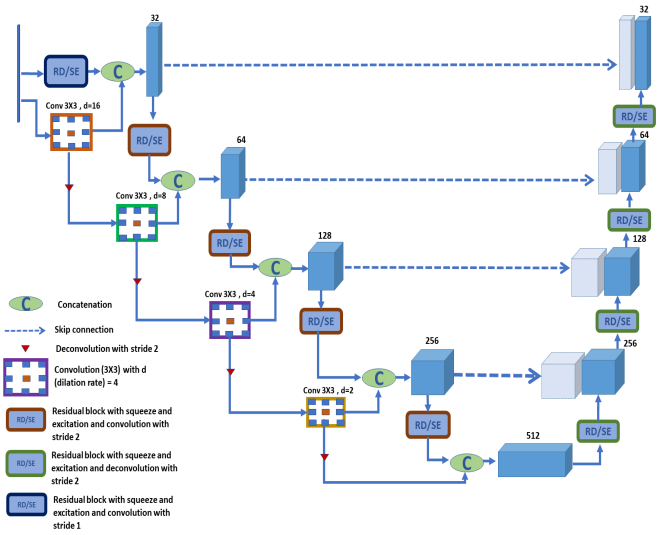

Fig. 2. The architecture of the proposed ResDUnet.

II. Methodology

Inspired by the success of U-net [10] and deep residual learning [11], we have chose to combine the strengths of both in one, seen in figure 2. It has been shown that deeper networks provide better performance [11]. However, training a deep U-net architecture is hard, especially for a limited number of training data, which can be defeated by employing a pre-trained network with fine tuning [12]. Therefore, He et al. [11] introduced residual blocks, which share the same idea of concatenating the input (identity short-cut) and propagating the low fine details, which enhances network performance without the need for going deeper. Accordingly, we chose to replace the basic U-net units, seen in figure 3(a) with residual blocks, seen in figure 3(c). Those residual blocks will further contribute in feature propagation at both, the encoding and decoding path. Also, propagating fine feature details to all coarser layer is done through convolution with stride instead of pooling (which is originally used in U-net), this convolution process, enhances features reuse, without the need of sophisticated deeper architecture. Each residual unit is defined as:

$$\begin{aligned} y_i &= h(x_i) + F(x_i, W_i) \\ x_{i+1} &= f(y_i) \end{aligned} \tag{1}$$

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. According to He et al. [13], the best combination is achieved by the pre-activation design illustrated in figure 3(c), which was also employed in our model.

Within each residual block, squeeze and excitation (SE) unit is added before identity mapping. According to Hu et al. [14], integrating SE blocks into convolution neural networks help boost the overall network performance, specifically when added to residual and inception networks. The mechanism of SE blocks allows the network at earlier layers

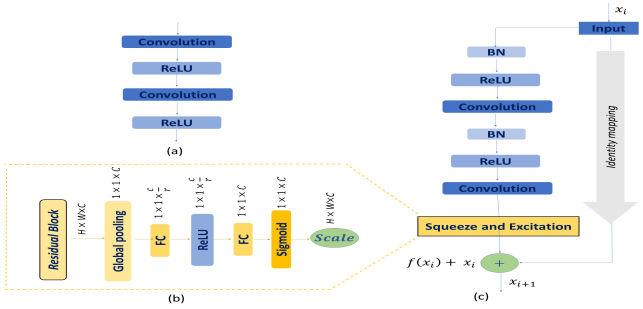

Fig. 3. Building blocks. (a) U-net basic building block. (b) Residual block with squeeze and excitation added before identity mapping. (c) Squeeze and excitation unit

to excites informative features, and strengths the shared low-level representation, while at later layers, it responds to different input in a highly class-specific manner, more details explanation can be seen in [14].

The structure of SE blocks is shown in figure 3(b). Each block 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 operation is done through global average pooling to generate channel wise statistics, which embed global distribution of channel-wise feature responses (C), allowing information from the global receptive field of the network to be used by all its layers. Second, the aggregated channels are passed to an excitation operation which is a simple self gating mechanism using sigmoid activation function. Precisely, excitation is added to fully capture channel wise dependencies and learn non-linear and non mutually exclusive relationship between channels. After global average pooling, there is a bottleneck with two fully convolution layer (FC) at a reduction ratio (r). Finally, the output of the whole SE block is obtained by rescaling the transformed feature maps with the activations, which acts as weights adapted to the input features.

Although U-net provides a notable performance through skip connections. This direct concatenation process ignores the contribution of all semantic strengths in the segmentation procedure, since it does not take into consideration different scales of the feature map [10], [12]. Therefore, we chose to tackle multi-scale feature extraction by adopting cascaded dilated convolution [15] with different dilation rates across the encoding path.

In comparison with conventional convolution, dilated convolution is able to achieve a larger receptive field size without increasing the number of kernel parameters. Dilated convolution expands the alignments of kernel weight with a dilation rate (d), increasing this factor provides more sparse weights, so the size of the kernel increases accordingly. Increasing the size of the kernel will in turn enlarge the field of view of filter to incorporate global and multi-scale context. According to [15], dilated convolution is applied over the

two-dimensional feature map x, where for each location i on the output y, a filter w is applied as shown in:

$$y[i] = \sum_k x[i + d.k]w[k] \quad (2)$$

where the dilation rate d, is equivalent to the stride with which the input signal is sampled. This process resembles convolving the input x with the up-sampled filters produced by inserting (d-1) zeros between two consecutive filter values along each spatial dimension. Thereby, using dilated convolution rates, we can adjust the filter's field of view, to capture multi-context information.

A segmentation mask generated from simply one dilation rate throughout the whole segmentation process still does not cover all semantic strengths [15], [16]. Therefore, in order to adapt to the resolution change along the encoding path and at the same time tackle the problem of left ventricle shape and size variability, we chose the methodology of cascaded dilated convolution with different rates (seen in figure 2, along the encoding path). This cascading mechanism is adopted to cover all scales in the image and further propagate this information along skip connections to the corresponding high level semantic.

III. Material and Implementation details

In this work, we use the public available dataset, CAMUS (Cardiac Acquisition for Multi-structure Ultrasound Segmentation) [8]. The dataset contains 2D echocardiographic for 500 patients, annotated by two different experts. For each patient, two views were captured, apical four chamber view (A4C) and apical two chamber view (A2C). For each view, two frames were taken, end diastole and end systole frame.

Images are resized down to a resolution of 256×256. Each residual block consists of two convolution layers with a kernel size of (3× 3), each layer is followed by batch normalization and ReLU function. Following the last batch normalization within each residual, a squeeze and excitation block is added. The reduction ratio for squeeze and excitation unit is set to 8, which provides the lowest overall error with Resnet, inspired by [14]. Moreover, Max pooling layers throughout the encoding path are replaced by deconvolution with a stride of 2.

Given that the feature map size is reduced to half as we go down the encoding path, we followed the same tuning of dilation rates used in [17], but in an inverted way to map the resolution of each feature map. We started the first dilated convolution with an dilation rate of 16 and decreased the rate with a factor of 2 till we reach the bottom layer (where the spatial information is most compressed) and used dilation rate of 2, which maps to a standard convolution.

The performance of proposed model was validated using 10 fold cross validation strategy, where each fold contains 50 patients. For each of the 10 test sets, the remaining 9 folds (450 patients), were divided into 8 folds for training and one fold (50 patients) for validation.

The model was implemented in Python environment with keras and Tensorflow. For optimization, Adam optimiser was used at 1000 epochs, with a learning rate of 0.0001. For the network objective function, Binary cross entropy with weight decay was used. All computation were carried out using Nvidia GeForce with RTX 2070 GPU.

TABLE I

Statistical evaluation for ResDUnet in comparison with other segmentation models (mean value ± standard deviation)

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	$\mathbf{0.951 \pm 0.030}$	$\mathbf{4.5 \pm 1.2}$	$\mathbf{1.4 \pm 1.2}$

IV. Evaluation results and discussion

To evaluate the performance of the proposed model we have utilized three standard segmentation performance metrics: Dice similarity coefficient (DSC), Hausdorff distance (HD), and mean absolute distance (MAD).

Segmentation performance of the proposed model is illustrated in table I in comparison with U-net [8] and deeplabv3 [15]. It can be seen that our model has outperformed the performance of deeplabv3 and U-net by 8.4% and 1.2%, respectively. Also table II, shows the dice similarity measure achieved by adding residual block with squeeze and excitation, which also outperformed the basic U-net model, and this result got boosted by 0.9% 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 of the left ventricle.

Figure 4, shows a qualitative comparison between the predicted contour from the proposed model versus U-net and deeplabv3, where expert manual segmentation is depicted in white. At the best case, where the left ventricular wall is well defined (good quality image), seen in figure 4(a), the three models provide comparable segmentation in reference to manual delineation, with a better performance seen in the proposed model. In the case of fair quality image, seen in figure 4(b), the two models failed to precisely constrain the left ventricle in comparison with our proposed work, which outperformed in that case, despite the leakage of left ventricle region out of the cone region. Finally, in the poor quality image, shown in figure 4(c), where the ventricular wall is very ill defined, deeplabv3 provided a poor segmentation, while our proposed model provided comparable results with U-net, but with a slight efficiency in segmentation contour.

Better results are achieved by U-net when compared with deeplabv3 due to the concatenation of low level features with the corresponding high level features at the decoding path. While, the proposed model has outperformed the results achieved by U-net due to the incorporation of cascaded dilated convolution which help capture multi-scale features in the left ventricle and propagate those features along both, skip connections and encoding path. Also, the presence of the squeeze and excitation block integrated within each residual block has enriched the feature extraction process with more informative features and minimized less informative ones.

TABLE II

SEGMENTATION RESULTS AFTER ADDING EACH MODULE; RD:RESIDUAL BLOCKS; SE:SQUEEZE AND EXCITATION; CD:CASCADED DILATION

Method	DSC	HD	MAD
RD/SE	0.942 ± 0.143	3.5 ± 1.9	2.2 ± 1.8
RD/SE/CD	$\mathbf{0.951 \pm 0.030}$	$\mathbf{4.5 \pm 1.2}$	$\mathbf{1.4 \pm 1.2}$

Due the large variability of the dataset in terms of quality and patients pathology, our model has partially failed in some cases where the left ventricular wall is very ill defined, as previously shown in figure 1(c), where parts of the wall were hard to visualize, which required an expert cardiologist. Also, in some other cases, the left ventricle was too large which was also challenging to accurately segment, and expert segmentation is leaked out from the cone region, as shown in figure 4(b). These properties already result in inter-observer and intra-observer variability. However, we have augmented the dataset with different transformations to ensure the generalization of the model.

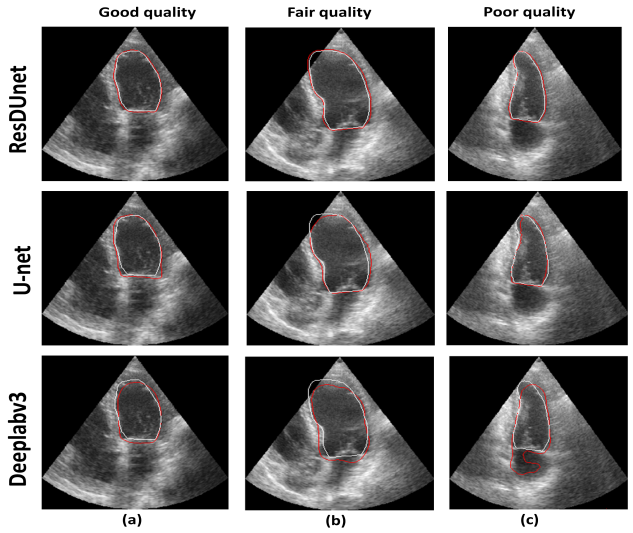

Fig. 4. LV segmentation by ResDUnet, U-net ,and Deeplabv3. Red contours present predicted segmentation. White contours present gold standard.(a) Apical-4 chamber at end systole phase. (b) Apical-4 chamber at end diastole phase. (c) Apical-2 chamber at end diastole phase. Image quality is shown above each column.

V. CONCLUSION

In this work, we have provided a simple yet robust deep learning model, which implies that, performance gains can be achieved when state-of-the-art U-net is integrated with innovative concepts such as residual, squeeze and excitation, and dilated convolution. Despite the challenging nature of echocardiographic images with ambiguous ventricular wall and variability of shape and size of the left ventricle, our model was able to outperform state-of-the-art methods. This out performance was achieved by capturing multi-scale semantic features through cascaded dilated convolutions, and utilizing residual blocks for extracting finer feature details without the need for deeper architecture. Also, squeeze and excitation blocks have further incorporated in providing more discriminative features through suppressing less important ones. For further analysis, we intend to investigate the performance of the proposed model on other different modalities.

REFERENCES

[1] Denisa Muraru, Luigi P Badano, Gianluca Piccoli, Pasquale Gianfagna, Lorenzo Del Mestre, Davide Ermacora, and Alessandro 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.

[2] Pamela S Douglas, Jeanne M DeCara, Richard B Devereux, Shelly Duckworth, Julius M Gardin, Wael A Jaber, Annitta J Morehead, Jae K Oh, Michael H Picard, Scott 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.

[3] Samaneh Mazaheri, Puteri Suhaiza Binti Sulaiman, Rahmita Wirza, Fatimah Khalid, Suhaini Kadiman, Mohd Zamrin Dimon, and Rohollah Moosavi Tayebi, "Echocardiography image segmentation: A survey," in *2013 International Conference on Advanced Computer Science Applications and Technologies*, 2013, pp. 327–332.

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

[5] Alison Noble and Djamal Boukerroui, "Ultrasound image segmentation: a survey," *IEEE Trans. on medical imaging*, vol. 25, no. 8, pp. 987–1010, 2006.

[6] Gustavo Carneiro, Jacinto C Nascimento, and Freitas, "The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods," *IEEE Trans. on Image Processing*, vol. 21, no. 3, pp. 968–982, 2011.

[7] Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Stuart A Cook, Antonio De Marvao, Timothy Dawes, Declan 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.

[8] Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas stvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al., "Deep learning for segmentation using an open large-scale dataset in 2d echocardiography," *IEEE Trans. on medical imaging*, 2019.

[9] Joao Pedrosa, Sandro Queirs, Olivier Bernard, Jan Engvall, Thor Edvardsen, Eike Nagel, and Jan 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.

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

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[12] JLaESaT Darrell and UC Berkeley, "Fully convolutional networks for semantic segmentation," *UC Berkeley*, 2014.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," 2016.

[14] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu, "Squeeze-and-excitation networks," 2017.

[15] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[16] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," 2016.

[17] J. Long, E.Shelhamer, and T.Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.