2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Simultaneous Right Ventricle End-diastolic and End-systolic Frame Identification and Landmark Detection on Echocardiography

[1]Zhaohui Wang, [1]Jun Shi, [1]Xiaoyu Hao, [1]Ke Wen, [1]Xu Jin, [1]Hong An

Abstract— **End-diastolic (ED) and end-systolic (ES) frame identification and landmark detection are crucial steps of estimating right ventricle function in clinic practice. However, the complex morphology of the right ventricle and low-quality echocardiography pose challenges to these tasks. This study proposes a multi-task learning (MTL) framework to simultaneously identify the right ventricle ED and ES frames and detect anatomical landmarks for echocardiography. The framework contains an encoder and two branches: frame-branch and landmark-branch. The convolution neural network (CNN) encoder is employed for extracting the shared features of two branches. The frame-branch is built with a recurrent neural network (RNN) to select ED and ES frames. A heatmap-based model is used as the landmark-branch to detect the landmarks. Furthermore, instead of directly regressing the indexes of ED/ES frames, we form the frame identification as a curve regression problem, which achieves considerable performance. Experiments performed on the echocardiography dataset of 105 patients validate the effectiveness of the proposed approach, which leads to the average frame difference of 1.59 ($\pm$1.34) frames (ED) and 1.56 ($\pm$1.35) frames (ES) on the frame identification task, and the percentage of correctly predicted landmarks is 83.3%. These results demonstrated that our method outperforms most existing methods.**

I. INTRODUCTION

Cardiovascular diseases (CVDs) are the leading cause of death globally. The evaluation of right ventricular (RV) function plays a critical role in predicting the morbidity and mortality of patients presenting with signs and symptoms of cardiopulmonary disease[1]. Benefits from its painless, noninvasive and being the only modality that able to image heart in real-time, echocardiography is one of the most commonly used modalities in clinic practice.

Right ventricular ejection fraction (RVEF) measures the efficiency of the heart to pump into the pulmonary circulation. Identifying the end-diastolic (ED)/end-systolic (ES) frames in the echocardiography sequence is the first step of quantifying RVEF. Then, localizing the anatomical landmarks of right ventricular structures can help accurately computing the end-diastolic volume (EDV) and end-systolic volume (ESV) [2], [3]. However, compared to the left ventricular (LV), it is much more difficult to assess the RV function due to its fuzzy boundaries and complex shape. Furthermore, the low-resolution and noise leading that ultrasound images have more challenges to analyze than other medical images.

Numerous efforts have been made to automatically identify the ED/ES frames and detect anatomical landmarks on echocardiography. For frame identification, Kachenoura et al. [4] proposed one of the earliest automatic methods. They combine the intensity variation curve in the mitral region and the correlation coefficients between the ED image and other images to identify the ES frame. The limit of this method is that it is not full-automatic and manual identification is required. Recently, deep-learning based methods show great success in medical image analysis. Kong et al. [5] integrated CNN with an RNN for full-automatically identify ED/ES frames from cine MRI. Then, Dezaki et al. [6] applied the method to estimate the ES and ED frames from an echocardiography cine series and proposed global extrema (GE) loss to improve the prediction performance. While methods mentioned above focus on LV, the RV quantification is of equal importance, which not enough work has been devoted to.

With regard to landmark detection problem, the framework of heatmap regression with full convolutional networks (FCN) [7], [8] achieved considerable results. Inspired by the work of human pose estimation [9], Payer et al. [10] considered the structure landmark detection problem as a heatmap regression problem, which achieved good performance on a limited number of datasets. Building upon the approach of [10], Zhong et al. [11] proposed a two-stage network to predict landmarks on x-ray images, which achieved state-of-the-art result on a public cephalometric X-ray dataset [12].

However, it is not efficient enough that designing and training two models for frame identification and landmark localization separately. Multi-task learning (MTL) can help not only improving model generalization ability but also saving memory and computation resources with shared features learning in only one network. Recently, various studies have been documented on MTL in cardiac image analysis. Xue et al. [13] proposed a deep MTL network to simultaneously predict all indices of the left ventricle. Xu et al. [14] also take the advantage of MTL to achieve view classification and landmark detection on Abdominal Ultrasound Images, which outperforms single-task models.

In this paper, we 1. Propose an MTL framework that can learn to identify ED/ES frames and detect anatomical landmarks simultaneously; 2. Design a curve regression strategy to improve the performance of the frame identification task; 3. Leverage heatmap-based model to localize anatomical landmarks of RV. Experiments validate the effectiveness and accuracy of the proposed approach on both ED/ES frame identification and landmark localization task

[1]School of Computer Science and Technology, University of Science and Technology of China, Hefei, China

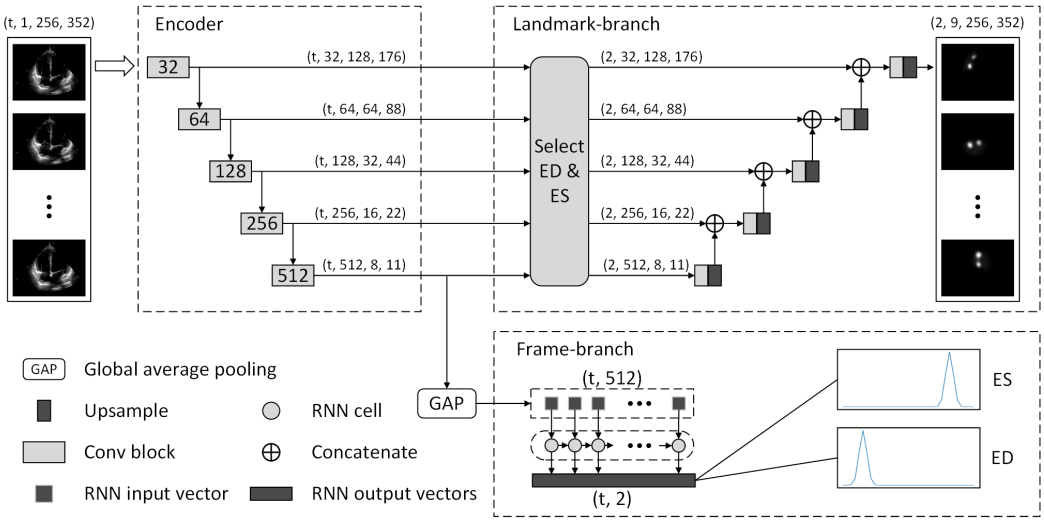

Fig. 1. The architecture of the proposed Multi-task learning framework

II. METHODOLOGY

As depicted in Fig.1, the proposed MTL framework contains an encoder and two branches: one for ED/ES frame identification (frame-branch) and the other for anatomical landmark detection (landmark-branch). The encoder is a convolution neural network (CNN) network learning the features shared by two branches. Frame-branch employs a recurrent neural network (RNN) network to predict the probability of each frame that is an ED or ES frame. Landmark-task is a decoder that upsamples the features to the size of the input image and generates the landmark heatmaps.

A. Frame Identification With Synthesized Target

To extract sufficient spatial and temporal information of echocardiography cine series, this paper follows the CNN + RNN framework employed in [5], [6]. The CNN encoder learns the spatial information of each frame while an RNN is used to extract the temporal information of the ultrasound sequence. Between the encoder and RNN is a global average pooling layer (GAP). The output of the frame-branch is two vectors (one for ED and the other for ES) of length t, where t is the number of frames in the input ultrasound sequence. For the training target of the frame-branch, this paper use Gaussian distribution to generate two curves which indicate the probability of each frame being an ED or ES frame (Fig.2 bottom).

During the training phase, the loss function is a mean squared error (MSE) loss:

$$L_{mse} = \sum_{n=1}^{N}\sum_{t=1}^{T} ||y^{(n,t)} - \hat{y}^{(n,t)}||^2 \tag{1}$$

Where $y^{(n,t)}$ and $\hat{y}^{(n,t)}$ are the ground truth and the output of tth frame in the nth sample.

While in the testing phase, the indexes of frames with a max value of ED/ES output are computed to indicate the predicted ED/ES frames:

$$\begin{aligned} Idx_{ed} &= \arg\max_{1\le t\le T} \hat{y}_{ed}^{(n,t)} \\ Idx_{es} &= \arg\max_{1\le t\le T} \hat{y}_{es}^{(n,t)} \end{aligned} \tag{2}$$

B. Heatmap-based Landmark Detection

This paper adopts heatmap-based method to build the landmark-branch. Rather than directly predict the coordinates of each landmark, the heatmap-based methods make the model output heatmaps that have the same size as the input image. In heatmaps, a pixel with higher intensity means that it is more likely to be the corresponding landmark. The framework proposed in this paper uses a CNN decoder as the landmark-branch to regressing the heatmaps (Fig.1 upper right). Only the features of the ED/ES frame are selected and fed to the landmark-branch. The target heatmap is generated for each anatomical structure with the 2-D Gaussian function.

To overcome the imbalance of background pixels and landmark pixels, we take the advantage of top-k cross-entropy [15] to train the landmark-branch, which is defined as:

$$L_{topk\_ce} = \frac{1}{k}\sum_{i=1}^{k} l_i(-\sum_{c=1}^{C} y_c \log \hat{y}_c) \tag{3}$$

Where $-\sum_{c=1}^{C} y_c \log \hat{y}_c$ is the cross entropy loss function over heatmaps, $l_i(\cdot)$ compute the kth largest value.

During testing, the ED/ES frames are selected according to the output of the frame-branch. The predicted landmark coordinates are determined by the position of pixels that have the maximum responses in corresponding heatmaps.

Finally, the loss function of the whole framework is computed as:

$$L_{total} = (1-\alpha)L_{mse} + \alpha L_{topk\_ce} \tag{4}$$

Where α is a hyper-parameter to control the weight of two loss terms and empirically set to 0.6 in this work.

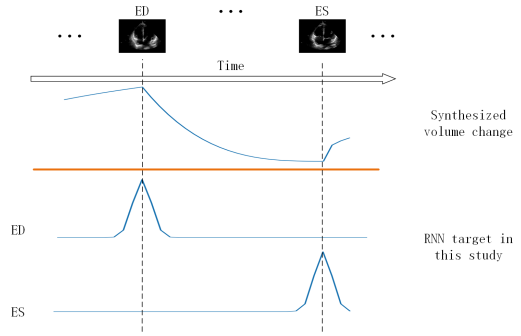

Fig. 2. Frame-branch prediction target

III. MATERIAL AND IMPLEMENTATION DETAILS

In this study, 1019 echo cine series of 105 participants are obtained from The First Affiliated Hospital of USTC, and no ethical approval is required. All of the subjects are adults with an average age of 48 and the male-to-female ratio is approximately 0.47. Among the participants, 32 were healthy while 73 with pulmonary hypertension or other heart diseases. The whole dataset is divided into 3 mutually exclusive sets: 60% for training, 20% for validation, and 20% for testing.

We employ Resnet-50 [16] as the encoder, stacked convolutional blocks as heatmap decoder and bidirectional GRU [17] as frame-task branch. Note that any other CNN/RNN network can replace them depending on the task and deployment environment. For the landmark task, there are 7 structures (RV endocardium, RV septum, tricuspid annulus, conal septum, pulmonic valve, RV apex, septal edge) to be predicted. The network is trained by Adam [18] optimizer with a learning rate of 0.001 and batch size of 4. All experiments are run on devices with CPU: Intel Xeon E5-2695 v4 @ 2.10GHz, GPU: Nvidia Tesla V100 (16GB).

IV. EVALUATION RESULTS AND DISCUSSION

A. metrics

Following the convention, average frame difference (aFD) is computed as the metric to evaluate the frame identification task. Assuming that the predicted ED/ES frame index of sequence i is $\hat{I}_i$, and the ground-truth is I_i, then aFD is defined as:

$$aFD = \frac{1}{N}\sum_{i=1}^{N}|I_i - \hat{I}_i| \tag{5}$$

The error of landmark localization is estimated with the point-to-point error (PE). PE evaluates the Euclidean distance between the target coordinate x_i and the predicted coordinate $\hat{x}_i$ of each landmark L_i:

$$PE_i = ||x_i - \hat{x}_i||_2 \tag{6}$$

With the incompletely labeled landmarks, we also leverage the number of identifications correct (ID) ([19]) to estimate landmark detection performance. A landmark is predicted correctly if the closest target landmark is correct and the distance between the predicted point and target one is less than 20 mm. Then the percentage of correctly identified landmarks over all landmarks (ID_{rate}) is used as the localization criterion.

B. results

The experiments are tested with 5-folds cross-validation. Tab.I shows the comparison of the proposed method and other methods in ED/ES frame identification. It is observed that the proposed framework achieves aFD of 1.59 ($\pm$1.34) frames on ED identifying and 1.56 ($\pm$1.35) frames on ES identifying, which means that our method works better on right ventricle echocardiography.

TABLE I
EVALUATION OF ED/ES FRAME IDENTIFICATION (AFD)

Method	ED (Mean $\pm$ std)	ES (Mean $\pm$ std)
Kong et al. [5]	1.83 $\pm$ 1.46	1.65 $\pm$ 1.32
Dezaki et al. [6]	2.05 $\pm$ 1.54	1.79 $\pm$ 1.43
Proposed	**1.59 $\pm$ 1.34**	**1.56 $\pm$ 1.35**

Fig.3 shows the representative samples of landmark detection. The statistical comparison of our landmark-branch and other encoder-decoder network benchmarks can be seen in Tab.II. The presented MTL framework achieves ID_{rate} of 0.833, and the localization error of our method for ED frame is 5.64 ($\pm$8.01) mm and for ES frame is 5.65 ($\pm$7.60) mm that proves the effectiveness of the proposed framework.

TABLE II
LANDMARK DETECTION RESULTS

	ID_{rate}	PE (Mean $\pm$ std)	
		ED (in mm)	ES (in mm)
Unet [8]	0.829	5.97 $\pm$ 7.72	5.92 $\pm$ 7.46
Proposed (STL)	0.828	5.95 $\pm$ 7.54	5.90 $\pm$ 7.09
Proposed (MTL)	**0.833**	**5.64 $\pm$ 8.01**	**5.65 $\pm$ 7.60**

C. discussion

The effectiveness of the proposed method is obtained from three aspects. Firstly, the curve regression strategy is adopted to handle the complexity of RV for the frame identification task. previous work ([5], [6]) synthesized a ventricle volume variation curve as the regression target (Fig.2 upper), which does not work well in our right ventricle dataset owing to the complex RV anatomy. This paper generates two curves to directly predict the probability of each frame being ED or ES which is more appropriate in the RV situation. Secondly, the top-k cross-entropy loss is leveraged to solve the pixel imbalance problem. Most pixels of the target heatmaps are background, resulting in the model being more likely to classify a pixel as background. The top-k loss could let the network pay more attention to the foreground pixels. Finally, the MTL framework helps the network trained faster. Despite two branches learn different tasks, they extract common spatial and temporal information from the same input sequence.

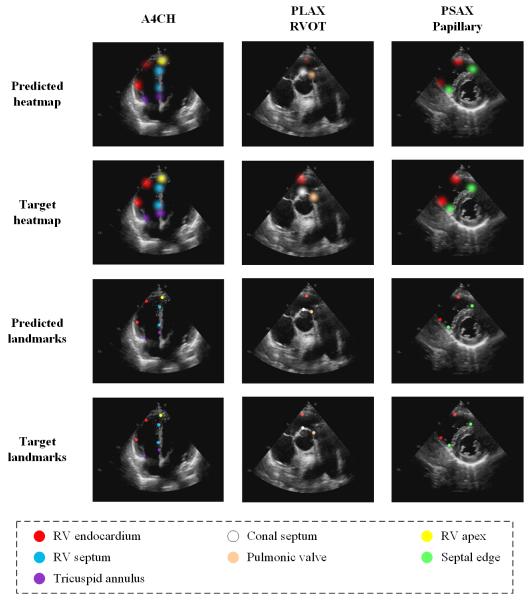

Fig. 3. Presentative samples of landmark detection on various views

With the MTL framework, the shared encoder structure can help two branches learning common hidden features and accelerate the convergence of the network.

V. CONCLUSIONS

This study proposes a joint learning framework that simultaneously achieves the automatic ED and ES frame identification and anatomical landmark localization for echocardiography. In the proposed method, we adopt the strategy of curve regression to improve the performance of frame identification. Besides, a heatmap-based model is designed to detect the landmarks. The effectiveness of our method on right ventricle echocardiography is validated by experiments. In addition, the proposed framework is flexible and the components could be replaced by any other CNN/RNN models depending on the specific task and deploy environment.

REFERENCES

[1] Roberto M Lang, Luigi P Badano, Victor Mor-Avi, Jonathan Afilalo, Anderson Armstrong, Laura Ernande, Frank A Flachskampf, Elyse Foster, Steven A Goldstein, Tatiana Kuznetsova, et al., "Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging," *European Heart Journal-Cardiovascular Imaging*, vol. 16, no. 3, pp. 233–271, 2015.

[2] Florence H Sheehan, Philip J Kilner, David J Sahn, G Wesley Vick III, Karen K Stout, Shuping Ge, Willem A Helbing, Mark Lewin, Alan J Shurman, Emanuela Valsangiacomo Buechel, et al., "Accuracy of knowledge-based reconstruction for measurement of right ventricular volume and function in patients with tetralogy of fallot," *The American journal of cardiology*, vol. 105, no. 7, pp. 993–999, 2010.

[3] Nicole M Bhave, Amit R Patel, Lynn Weinert, Megan Yamat, Benjamin H Freed, Victor Mor-Avi, Mardi Gomberg-Maitland, and Roberto M Lang, "Three-dimensional modeling of the right ventricle from two-dimensional transthoracic echocardiographic images: utility of knowledge-based reconstruction in pulmonary arterial hypertension," *Journal of the American Society of Echocardiography*, vol. 26, no. 8, pp. 860–867, 2013.

[4] Nadjia Kachenoura, Annie Delouche, Alain Herment, Frédérique Frouin, and Benoît Diebold, "Automatic detection of end systole within a sequence of left ventricular echocardiographic images using autocorrelation and mitral valve motion detection," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 4504–4507.

[5] Bin Kong, Yiqiang Zhan, Min Shin, Thomas Denny, and Shaoting Zhang, "Recognizing end-diastole and end-systole frames via deep temporal regression network," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 264–272.

[6] Fatemeh Taheri Dezaki, Zhibin Liao, Christina Luong, Hany Girgis, Neeraj Dhungel, Amir H Abdi, Delaram Behnami, Ken Gin, Robert Rohling, Purang Abolmaesumi, et al., "Cardiac phase detection in echocardiograms with densely gated recurrent neural networks and global extrema loss," *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1821–1832, 2018.

[7] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[8] 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.

[9] Tomas Pfister, James Charles, and Andrew Zisserman, "Flowing convnets for human pose estimation in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1913–1921.

[10] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler, "Regressing heatmaps for multiple landmark localization using cnns," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 230–238.

[11] Zhusi Zhong, Jie Li, Zhenxi Zhang, Zhicheng Jiao, and Xinbo Gao, "An attention-guided deep regression model for landmark detection in cephalograms," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 540–548.

[12] Ching-Wei Wang, Cheng-Ta Huang, Meng-Che Hsieh, Chung-Hsing Li, Sheng-Wei Chang, Wei-Cheng Li, Rémy Vandaele, Raphaël Marée, Sébastien Jodogne, Pierre Geurts, et al., "Evaluation and comparison of anatomical landmark detection methods for cephalometric x-ray images: a grand challenge," *IEEE transactions on medical imaging*, vol. 34, no. 9, pp. 1890–1900, 2015.

[13] Wufeng Xue, Gary Brahm, Sachin Pandey, Stephanie Leung, and Shuo Li, "Full left ventricle quantification via deep multitask relationships learning," *Medical image analysis*, vol. 43, pp. 54–65, 2018.

[14] Zhoubing Xu, Yuankai Huo, JinHyeong Park, Bennett Landman, Andy Milkowski, Sasa Grbic, and Shaohua Zhou, "Less is more: Simultaneous view classification and landmark detection for abdominal ultrasound images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 711–719.

[15] Yanbo Fan, Siwei Lyu, Yiming Ying, and Bao-Gang Hu, "Learning with average top-k loss," *arXiv preprint arXiv:1705.08826*, 2017.

[16] 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.

[17] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[18] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] Ben Glocker, Johannes Feulner, Antonio Criminisi, David R Haynor, and Ender Konukoglu, "Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2012, pp. 590–598.

[20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.