Recurrent Aggregation Learning for Multi-view Echocardiographic Sequences Segmentation

Ming Li[1,2], Weiwei Zhang[1], Guang Yang[3,4], Chengjia Wang[5], Heye Zhang[6(✉)], Huafeng Liu[7], Wei Zheng[1(✉)], and Shuo Li[8]

[1] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
zhengwei@siat.ac.cn
[2] Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, China
[3] Cardiovascular Research Centre, Royal Brompton Hospital, London SW3 6NP, UK
[4] National Heart & Lung Institute, Imperial College London, London SW7 2AZ, UK
[5] BHF Centre for Cardiovascular Science, University of Edinburgh, Edinburgh, UK
[6] School of Biomedical Engineering, Sun Yat-Sen University, Shenzhen, China
zhangheye@mail.sysu.edu.cn
[7] Zhejiang University, Hangzhou, China
[8] Western university, London, ON, Canada

Abstract. Multi-view echocardiographic sequences segmentation is crucial for clinical diagnosis. However, this task is challenging due to limited labeled data, huge noise, and large gaps across views. Here we propose a recurrent aggregation learning method to tackle this challenging task. By pyramid ConvBlocks, multi-level and multi-scale features are extracted efficiently. Hierarchical ConvLSTMs next fuse these features and capture spatial-temporal information in multi-level and multi-scale space. We further introduce a double-branch aggregation mechanism for segmentation and classification which are mutually promoted by deep aggregation of multi-level and multi-scale features. The segmentation branch provides information to guide the classification while the classification branch affords multi-view regularization to refine segmentations and further lessen gaps across views. Our method is built as an end-to-end framework for segmentation and classification. Adequate experiments on our multi-view dataset (9000 labeled images) and the CAMUS dataset (1800 labeled images) corroborate that our method achieves not only superior segmentation and classification accuracy but also prominent temporal stability.

1 Introduction

Multi-view echocardiographic sequences delineation provides important insight for clinical diagnosis. The knowledge pattern of cardiac structures and textures associated with deforming tissues can be observed in echocardiographic sequence

D. Shen et al. (Eds.): MICCAI 2019, LNCS 11765, pp. 678–686, 2019.
https://doi.org/10.1007/978-3-030-32245-8_75

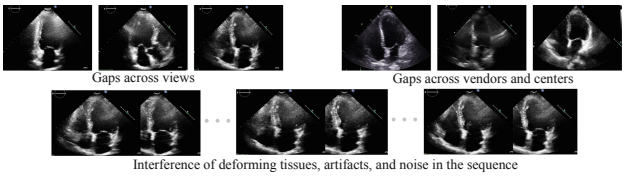

Fig. 1. Top left: multi-view samples (A2C, A3C, and A4C). Top right: A4C samples across vendors and centers. Bottom row: echocardiographic sequence

while in single frames the information is always missing and incomplete [1]. Echocardiographic sequence also permits the assessment of wall motion and identification of end-diastolic (ED) and end-systolic (ES) phases. Cardiologists usually check multi-view echocardiographic sequences in clinical decision-making [2]. The apical-2-chamber view (A2C), A3C, and A4C are the most commonly used views for the left ventricle (LV) functional assessment. Most clinical indexes of the LV (e.g., area, volume, and ejection fraction) are basically measured in these standard apical views. Segmentation of the LV is generally a prerequisite for such quantitative analysis [3]. In clinical routine, quantitative analysis of the LV still involves careful review and massive manual interpretation by experts, which is a tedious and time-consuming task. Thus, automatic methods are desired to facilitate this process. However, multi-view echocardiographic sequences segmentation remains a challenging task as illustrated in Fig. 1. First, the fuzzy border, huge noise, and abounding artifacts of echocardiographic images result in local missing and incomplete of the anatomical structures; Second, multi-view heterogeneous data varies in the anatomical structure, and image properties differ widely across vendors and centers; Third, in the sequence, artifacts and noise are much severer, and the motion of mitral valve, trabeculation, and papillary muscles also poses additional interference; Finally, limited labeled data restricts the performance of supervised learning based methods.

The application scenario of existing methods is always limited and only suitable under a specific situation. They mostly focus on specific view [4] or single frames (i.e., without considering the sequence) [5] or one single vendor and center [6]. As for sequence segmentation, existing methods try to leverage temporal information by using a deformable model combined with the optical flow [7,8] or fine-tuning pretrained CNN dynamically with first frame's label till the last frame [9]. The major downsides of these temporal methods are that they are computational cumbersome and not an end-to-end manner. The limited labeled data and specific application scenario confine the performance of existing methods and lead to the suboptimal solution.

To achieve a unified model for multi-view echocardiographic sequences segmentation, we propose a recurrent aggregation learning method (RAL). The workflow is depicted in Fig. 2. Pyramid ConvBlocks joint hierarchical ConvLSTMs are utilized to capture multi-level and multi-scale spatial-temporal information, enabling RAL the ability to harness the knowledge across heteroge-

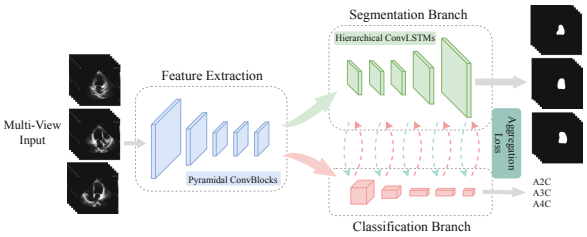

Fig. 2. Workflow overview of our method.

neous data (multi-view, multi-center, and multi-vendor). We further introduce a double-branch aggregation mechanism for segmentation and classification to lessen gaps across multi-view data. Different from existing methods, RAL fully exploits the long term spatial-temporal information in an end-to-end manner and does not depend on any deformable model or optical flow or pretrained segmentation models. RAL can accommodate heterogeneous data, not only generate accuracy segmentation results but also achieve the classification of different views at the same time and gain prominent temporal stability.

2 Method

RAL is built as an end-to-end framework and comprised of three key components: the feature extraction module, the segmentation branch, and the classification branch (as depicted in Fig. 2). The feature extraction module consists of pyramid dilated dense convolution blocks (ConvBlocks). The segmentation branch contains hierarchical recurrent architecture of multiple ConvLSTMs [10]. While the classification branch involves a series of aggregation downsample and fully connected layers.

Multi-level and Multi-scale Features Extraction. We design pyramid ConvBlocks architecture in the feature extraction module, which includes 5 ConvBlocks to extract multi-level and multi-scale features. Multi-level information provides the global geometric characteristic of the LV, while multi-scale information can help to strengthen thin and small regions, further refine the boundaries of the LV. They contribute to lessening the gap across views, vendors, and centers, increasing robustness to images conditions and the anatomical structure variations. One ConvBlock contains L densely connected dilated convolution layers as shown in Fig. 3, which can expand the receptive field and meanwhile preserve the resolution of feature maps. While the transition layer changes channels and resolution of feature maps by convolution and pooling. The feedforward information propagation from preceding l layers to $(l+1)^{th}$ layer can be formulated as

$$y_l = D(C(y_1, y_2, ..., y_{l-1})) \tag{1}$$

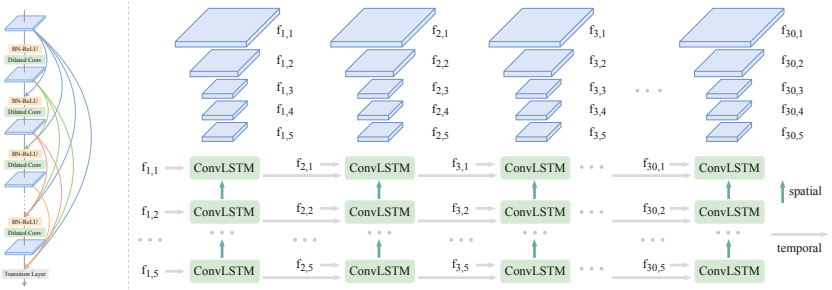

Fig. 3. Left: Dilated dense convolution block. Right: Hierarchical ConvLSTMs for spatial-temporal modeling.

where y_l are the output of the l^{th} layer, $C(\cdot)$ refers to the concatenation of previous layers' outputs. $D(\cdot)$ is a composite function of three connected operations: batch normalization (BN), rectified linear unit (ReLU), and dilated convolution. Five ConvBlocks generate multi-level and multi-scale features $f_t = \{f_{t,1}, f_{t,2}, f_{t,3}, f_{t,4}, f_{t,5}\}$ for frame t in the sequence.

Pyramid ConvBlocks endow RAL with the superior feature extraction ability and the LV region detection capacity in multi-level and multi-scale space, further contribute to capturing the global geometric characteristic of the LV and then establishing uniform semantic features. Thus RAL can detect and extract the LV accurately and robustly from not only ED and ES frames but also other frames in the sequence where the boundary is not clear (disturbed by noise and other tissues, see sequence samples in Fig. 1).

Recurrent Features Fusion for Spatial-Temporal Modeling. For sequence segmentation, capturing the LV characteristic over time is essential for temporal stability. Recent studies based on LSTM have shown great ability to learn sequential information. Inspired by [11,12], we conduct hierarchical ConvLSTMs to exploit long term spatial-temporal modeling as depicted in Fig. 3. We add recurrence in the temporal domain to generate prediction S_t for frame t in the sequence, which carries forward the LV information from previous frames to following frames and allows the matching between consecutive frames naturally. Additionally, we also add recurrence in the spatial domain for multi-level and multi-scale features fusion, which helps to integrate multi-level and multi-scale features efficiently.

The output $y_{t,k}$ of the k^{th} ConvLSTM at frame t depends on the following variables: (1) k^{th} level and scale feature $f_{t,k}$ from the feature extraction module; (2) the output $y_{t,k-1}$ of preceding $(k-1)^{th}$ ConvLSTM at the same frame t; (3) the output $y_{t-1,k}$ from the k^{th} ConvLSTM of previous frame $t-1$; (4) the hidden state representation $h_{t,k-1}$ from preceding $(k-1)^{th}$ ConvLSTM at the same frame t, which is the spatial hidden state; (5) the hidden state representation $h_{t-1,k}$ from the k^{th} ConvLSTM of previous frame $t-1$, which is the temporal hidden state. The information flow can be formulated as

$$x_{input} = [f_{t,k} \mid B(y_{t,k-1}) \mid y_{t-1,k}] \tag{2}$$

$$h_{state} = [h_{t,k-1} \mid h_{t-1,k}] \tag{3}$$

$$y_{t,k} = ConvLSTM_k(x_{input}, h_{state}) \tag{4}$$

where $B(\cdot)$ is the bilinear upsampling operator. At each time step, every ConvLSTM accepts hidden states and encoded spatial-temporal features from previous ConvLSTMs and frame, the corresponding extracted feature from the feature extraction module, it then outputs encoded spatial-temporal features to next ConvLSTM and frame. Finally, predictions S_t are generated by the last ConvLSTM at every frame.

Double-Branch Aggregation Learning. To further lessen the gaps across multi-view and refine multi-view segmentation results, we introduce a double-branch aggregation mechanism for simultaneous segmentation and classification of multi-view echocardiographic sequences as depicted in Fig. 2. Feature from the last ConvBlock is sent to the classification branch. Next, it goes through successive convolution and pooling operators to deeply aggregate with multi-level and multi-scale spatial-temporal features from the segmentation branch. Finally, the classification result is produced by fully connected layers.

Table 1. Specifications of our dataset (left) and the CAMUS dataset (right).

Vendor	Machines	Patients	Sequences	Images		A2C	A3C	A4C
Philips EPIQ 7C	60	180	5400	**Sequences**	100	100	100	
GE VIVID E9	20	60	1800	Training		240		
Philips IE33	20	60	1800	Testing		60		
Total	100	300	**9000**	Total		300		

CAMUS	A2C	A4C	Vendor Machine
Images	900	900	
Training	1600		GE VIVID E95
Testing	200		
Total	**1800**		

The segmentation branch generates multi-view segmentations while the classification branch discriminates the specific view. They are mutually promoted by deep aggregation of multi-level and multi-scale spatial-temporal features. The segmentation branch provides multi-level and multi-scale spatial-temporal information to guide the classification while the classification branch affords multi-view discriminative regularization to refine the segmentation results and further lessen the gaps across views. This double-branch aggregation mechanism endows RAL outstanding ability to adapt complex variations of anatomical structure.

Additionally, we propose an aggregation loss to dynamically facilitate the communication between the segmentation branch and the classification branch as illustrated in Fig. 2. The aggregation loss comprises the segmentation loss and classification loss. The segmentation loss is a combination of binary cross-entropy loss and dice loss. While the classification loss is categorical cross-entropy loss. Thus the aggregation loss function can be formulated as

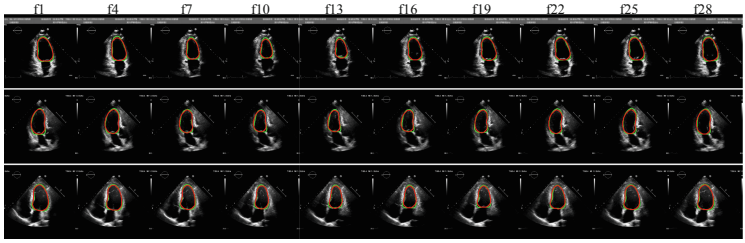

Fig. 4. The LV contours of multi-view sequences segmented by our method (red) and experts (green). Ten frames are selected from every sequence to fit the layout view. (top row: A2C; middle row: A3C; bottom row: A4C)

Table 2. Ablation results of our methods under different configurations.

Configurations	Accuracy	Dice	HD (mm)	MAD (mm)	Classification
Full	0.987 ± 0.005	0.919 ± 0.040	5.87 ± 3.46	2.90 ± 1.49	0.933
w/o classification	0.971 ± 0.008	0.910 ± 0.049	5.99 ± 3.69	3.10 ± 1.66	–
w/o ConvBlock	0.963 ± 0.015	0.907 ± 0.057	6.21 ± 4.95	3.27 ± 1.85	0.867
w/o temporal	0.955 ± 0.019	0.896 ± 0.062	6.64 ± 5.04	3.51 ± 1.93	0.917
5 w/o spatial	0.968 ± 0.011	0.911 ± 0.054	6.03 ± 4.16	3.08 ± 1.71	0.883

$$L_{segmentation} = -[G \cdot log(P) + (1 - G) \cdot log(1 - P)] + \frac{2 \cdot G \cdot P}{G + P} \tag{5}$$

$$L_{classification} = -\sum_{i=1}^{3} g_i \cdot log(p_i) \tag{6}$$

$$L_{aggregation} = \lambda_s \cdot L_{segmentation} + \lambda_c \cdot L_{classification} \tag{7}$$

where G and P denote ground truth and prediction of segmentation respectively, g and p refer to ground truth and prediction of classification separately, i indicates the type of view. Besides, λ_s and λ_c are the corresponding balance coefficients, both are chosen empirically during the training process.

3 Experiments

Datasets. To validate the efficiency of RAL, we built a large multi-view echocardiographic sequences dataset, which was acquired from three centers' various vendor machines (The Second People's Hospital of Shenzhen, The Third People's Hospital of Shenzhen, and Peking University First Hospital). We further evaluate RAL on the public CAMUS dataset [6]. Our dataset contains 300 sequences from 3 views and every sequence includes 30 frames. All 9000 frames were labeled by two experts. Figure 4 presents A2C, A3C, and A4C sequences samples segmented by RAL and experts. While the CAMUS dataset only contains manual

Table 3. Geometrical comparison results on our multi-view echocardiographic sequences dataset.

Methods	Accuracy	Dice	HD (mm)	MAD (mm)
RAL	0.987 ± 0.005	0.919 ± 0.040	5.87 ± 3.46	2.90 ± 1.49
U-Net	0.942 ± 0.030	0.883 ± 0.068	8.94 ± 6.87	3.72 ± 1.87
ACNN	0.959 ± 0.013	0.893 ± 0.061	7.70 ± 6.58	3.40 ± 1.57
U-Net++	0.937 ± 0.032	0.880 ± 0.072	9.01 ± 7.14	3.86 ± 2.01

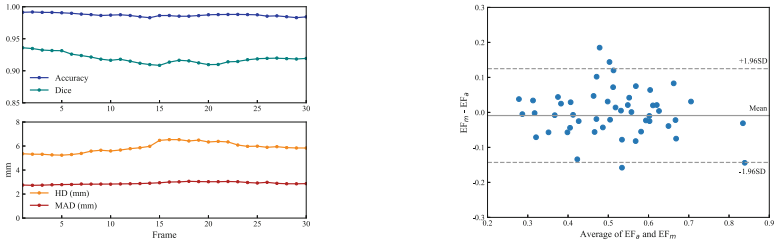

Fig. 5. Left: Mean of Accuracy, Dice, HD, and MAD at different frames of the cardiac cycle. Right: Bland-Altman analysis (EF_a and EF_m: ejection fraction calculated from automatic segmentations and manual labels)

labels at ED and ES frames, which was acquired from a single vendor and center. Table 1 shows the specifications of two datasets.

Evaluation Metrics. Accuracy, Dice, Mean Absolute Distance (MAD) and Hausdorff Distance (HD) are used to measure segmentation results. We further evaluate the segmentation performance with the ED, ES volume, and ejection fraction on the CAMUS dataset. We utilize the output of RAL to compute clinical indices according to standard guidelines [3].

Implementation Details. All images are resized to 256×256 for computational efficiency. We employ Adam with the learning rate of 0.001 as the optimizer. The dilated rates of 5 ConvBlocks are $1, 1, 2, 4, 8$ respectively, and every ConvBlock contains 6 layers. Besides, a dynamical decay mechanism is utilized to reduce the learning rate by monitoring the change of Dice. Ten-fold cross-validation was utilized to provide an unbiased estimation.

Ablation Study. We evaluate our method under different configurations to corroborate the necessity of every component in RAL. The classification branch, ConvBlock, spatial modeling and temporal modeling are removed respectively. Table 2 shows the ablation results, we can see that full RAL achieves higher mean values of Accuracy and Dice, lower mean values of HD and MAD, and lower standard deviations of all metrics compared against other configurations. RAL also achieves the best classification accuracy (0.933). Every single component brings important improvement for the LV segmentation, especially when adding recurrence in the temporal domain.

Table 4. Clinical comparison results on CAMUS dataset. (EDV: ED volume; ESV: ES volume; EF: ejection fraction; corr: Pearson correlation; mae: mean absolute error)

Methods	EDV			ESV			EF		
	corr	bias (ml)	mae (ml)	corr	bias (ml)	mae (ml)	corr	bias (%)	mae (%)
RAL	0.952	-7.5 ± 11.0	8.8	0.960	-3.8 ± 9.2	7.1	0.839	-0.9 ± 6.8	5.0
U-Net	0.954	-6.9 ± 11.8	9.8	0.964	-3.7 ± 9.0	6.8	0.823	-1.0 ± 7.1	5.3
ACNN	0.945	-6.7 ± 12.9	10.8	0.947	-4.0 ± 10.8	8.3	0.799	-0.8 ± 7.5	5.7
U-Net++	0.946	-11.4 ± 12.9	13.2	0.952	-5.7 ± 10.7	8.6	0.789	-1.8 ± 7.7	5.6

Comparison Study I: Geometrical. We compare RAL with U-Net, ACNN, and U-Net++ on our multi-view sequences dataset. As shown in Table 3, RAL outperforms other methods on all metrics, achieving the highest mean values of Accuracy (0.987) and Dice (0.919), the lowest mean values of HD (5.87 mm) and MAD (2.90 mm), and significantly lower standard deviations of all metrics. These strongly prove that RAL is able to accomplish the best region coverage, the highest contour accuracy, and the minimum distance error when processing multi-view echocardiographic sequences across multi-vendor and multi-center.

Comparison Study II: Clinical. We compare RAL with U-Net, ACNN, and U-Net++ on the CAMUS dataset to calculate clinical indices. As shown in Table 4, RAL obtained high correlation scores (0.952 for EDV, 0.960 for ESV, and 0.839 for EF), reasonably small biases and standard deviations, and relatively low mae (8.8 ml for EDV, 7.1 ml for ESV, and 5.0% for EF). Figure 5 presents a more intuitional result by Bland-Altman plot. 94% of the measurements locate in the ± 1.96 standard deviation in Bland-Altman plot. These results reveal the clinical potential of RAL.

Temporal Stability. We compute the mean of Accuracy, Dice, HD, and MAD at different frames of all echocardiographic sequences and then observe the volatility of each metric to assess the temporal stability. As shown in Fig. 5, RAL achieves stable mean values of all four metrics in the cardiac cycle, only exists moderate fluctuating in the middle of the sequence. This means spatial-temporal modeling of RAL is efficient. RAL achieves not only superior segmentation accuracy but also a good coherence of consecutive frames in the sequence.

Limitation. In Fig. 5, from ED to ES frames, we observe that Accuracy and Dice decay slightly while HD and MAD increase mildly, and all metrics keep relatively stable in the diastole but show feeblish recoverability. The sequential process carries errors forward resulting in accumulation of temporal errors in the cardiac cycle. Fortunately, the fluctuating rate is moderate and the worst results are still fairly good. This limitation could be alleviated via Bi-direction LSTM.

4 Conclusion

In this paper, we present a recurrent aggregation learning method to exploit long term spatial-temporal information for simultaneous segmentation and clas-

sification of multi-view echocardiographic sequences. Multi-level and multi-scale features are recurrently aggregated on both spatial domain and temporal domain for effective spatial-temporal modeling. A double-branch aggregation mechanism further brings multi-view discriminative regularization to refine the segmentation results. Adequate experiments of geometrical and clinical evaluation demonstrate that RAL achieves not only superior segmentation and classification accuracy, prominent temporal stability, but also high correlations on clinical indices.

Acknowledgment. This work is funded by the Shenzhen Basic Research Program (JCYJ20170818164343304, JCYJ20180507182432303).

References

1. Huang, X., et al.: Contour tracking in echocardiographic sequences via sparse representation and dictionary learning. Med. Image Anal. **18**(2), 253–271 (2014)
2. Madani, A., et al.: Fast and accurate view classification of echocardiograms using deep learning. NPJ Digital Med. **1**(1), 6 (2018)
3. Lang, R.M., 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. Eur. Hear. J.-Cardiovasc. Imaging **16**(3), 233–271 (2015)
4. Carneiro, G., et al.: The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods. IEEE Trans. Image Process. **21**(3), 968–982 (2012)
5. Chen, H., Zheng, Y., Park, J.-H., Heng, P.-A., Zhou, S.K.: Iterative multi-domain regularized deep learning for anatomical structure detection and segmentation from ultrasound images. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 487–495. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_56
6. Leclerc, S., et al.: Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE Trans. Med. Imaging (2019)
7. Pedrosa, J., et al.: Fast and fully automatic left ventricular segmentation and tracking in echocardiography using shape-based b-spline explicit active surfaces. IEEE Trans. Med. Imaging **36**(11), 2287–2296 (2017)
8. Zhang, N., et al.: Deep learning for diagnosis of chronic myocardial infarction on nonenhanced cardiac cine MRI. Radiology **291**(3), 606–617 (2019)
9. Yu, L., et al.: Segmentation of fetal left ventricle in echocardiographic sequences based on dynamic convolutional neural networks. IEEE Trans. Biomed. Eng. **64**(8), 1886–1895 (2017)
10. Xingjian, S., et al.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: NIPS, pp. 802–810 (2015)
11. Chen, J., et al.: Multiview two-task recursive attention model for left atrium and atrial scars segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 455–463. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_51
12. Yang, G., et al.: Multiview sequential learning and dilated residual learning for a fully automatic delineation of the left atrium and pulmonary veins from late gadolinium-enhanced cardiac MRI images. In: EMBC, pp. 1123–1127. IEEE (2018)