Published in final edited form as:

Med Image Comput Comput Assist Interv. 2021 ; 12901: 348–357. doi:10.1007/978-3-030-87193-2_33.

Multi-frame Attention Network for Left Ventricle Segmentation in 3D Echocardiography

Shawn S. Ahn¹, Kevinminh Ta¹, Stephanie Thorn², Jonathan Langdon⁴, Albert J. Sinusas^{2,3,4}, James S. Duncan^{1,3,4}

¹Department of Biomedical Engineering, Yale University, New Haven, CT, USA

²Section of Cardiovascular Medicine, Department of Internal Medicine, Yale University, New Haven, CT, USA

³Department of Electrical Engineering, Yale University, New Haven, CT, USA

⁴Department of Radiology and Biomedical Imaging, Yale University, New Haven, CT, USA

Abstract

Echocardiography is one of the main imaging modalities used to assess the cardiovascular health of patients. Among the many analyses performed on echocardiography, segmentation of left ventricle is crucial to quantify the clinical measurements like ejection fraction. However, segmentation of left ventricle in 3D echocardiography remains a challenging and tedious task. In this paper, we propose a multi-frame attention network to improve the performance of segmentation of left ventricle in 3D echocardiography. The multi-frame attention mechanism allows highly correlated spatiotemporal features in a sequence of images that come after a target image to be used to augment the performance of segmentation. Experimental results shown on 51 in vivo porcine 3D+time echocardiography images show that utilizing correlated spatiotemporal features significantly improves the performance of left ventricle segmentation when compared to other standard deep learning-based medical image segmentation models.

Keywords

3D echocardiography; Multi-frame attention; Segmentation

1 Introduction

Cardiovascular diseases (CVDs) are the number one cause of death globally with more than 17 million deaths each year [21]. CVDs include a wide range of disorders of the heart and vessels including coronary artery disease, valvular diseases, and cerebrovascular diseases. Echocardiography is one of the main imaging modalities used routinely by cardiologists to assess the functional status of the heart. 2D echocardiography is typically used to observe the motion of the heart during the cardiac cycle as well as to quantify the ejection fraction by measuring the size of the left ventricle cavity at end-diastole and end-systole. 3D

echocardiography has been gaining more utility in clinical practice, but measuring and segmenting the left ventricle cavity in 3D volume images are still challenging tasks that take considerable time even for experts.

Left ventricle segmentation in 3D echocardiography remains a difficult task due to intrinsic limitations such as low image resolution and high dimensionality. Also, there is a lack of ground truth labels for all time frames within a given 3D echocardiography sequence. Sonographers and cardiologists typically assess only the end-diastole (ED) and end-systole (ES) frames in clinical settings to analyze ejection fraction, but segmentation of left ventricle in all time frames is often critical to do further analysis in motion tracking and strain analysis [1,3,13,14,17,19].

Many recent deep learning algorithms have shown remarkable results with high expert-level accuracy. While these networks have shown promising results in Computed Tomography (CT) and Magnetic Resonance imaging (MRI), there is still a relative lack of progress in 3D cardiac segmentation in echocardiography [2]. Also, most of the previous neural network frameworks for cardiac segmentation tasks are based on single image inputs [4,5]. However, unlike CT and MR, echocardiography is inherently a spatiotemporal dataset. These make it difficult to transfer established frameworks for their use in images with low signal-to-noise ratio (SNR) like echocardiography. Although many works in left ventricle segmentation in 2D echocardiography have been done [6,8,12,18,26], full 3D volume characterization is still needed to accurately assess cardiac function using measurements like ejection fraction. Others have looked at non-deep learning approaches using spatiotemporal consistencies for left ventricle echocardiography segmentation [7]. However, they require manual tracings to initialize the sparse representations and appearance dictionaries, which makes it computationally expensive and slow. Previous works in computer vision and natural language processing communities have also looked at the use of temporal consistencies using attention mechanisms to focus in on regions of interest during the training of neural network models. Some examples of these attention mechanisms include global attention [25], local attention [11,25], and co-attention [10,22,24], which have all shown improvements in their respective domains. Attention mechanisms have also been shown to have improvement from standard neural network approaches in medical imaging segmentation [9,16].

Given the inherent spatiotemporal nature of an echocardiography image sequence, a notable characteristic of echocardiography is that there is interframe consistency among the images due to high temporal resolution. In addition, the myocardium is usually placed in the center of the image and is the only object that deforms throughout the time sequence. Therefore, we hypothesize that utilizing the spatiotemporal features with attention will improve the overall segmentation task performance in ultrasound. Thus, in this paper, we introduce a multi-frame attention architecture for segmentation of left ventricle in 3D echocardiography. The proposed attention module can be easily scalable to include multiple times frames to utilize spatiotemporal features in a time series medical image dataset. To our knowledge, this work is the first to successfully implement a spatiotemporal attention module using multiple time frames to segment the left ventricle in a volumetric echocardiography sequence.

2 Methods

2.1 Multi-Frame Attention Network for Segmentation

We propose to extract and utilize spatiotemporal features in multiple time frames to segment out the left ventricle myocardium in 3D+time echocardiography. Using multiple time frames allows us to take advantage of the high temporal resolution of echocardiography by looking at spatiotemporal consistencies among many time frames. The proposed multi-frame attention segmentation architecture is illustrated in Fig. 1. The idea of attention mechanism has been proposed by others previously [9,10,22] for use in segmenting out objects in a 2D video sequence. In this work, we adopt the idea of attention in a 3D+time dataset and expand its use by incorporating multiple time frames.

Given a target volume frame I_t , the volume frames that come immediately after within the same sequence will be selected, I_{t+i} , where $i = 1, \dots, n-1$, and n is the total number of frames in a given sequence of volumes. The network can adapt to incorporate as many time frames as hardware limitations allow. Figure 1 shows an instance where the inputs to the model are a target frame and 4 reference frames. This allows the target volume frame to utilize a series of 4 volume frames that follow to learn the spatiotemporal attention features. Each volume from a selected series will go through a standard encoder architecture with shared weights. For a 5-frame input, at the end of the encoder phase, there will be 5 feature maps, $F_t, F_{t+1}, F_{t+2}, F_{t+3}, F_{t+4}$. Each feature map corresponding to the input volume has the dimension of $c \times w \times h \times d$ where c = channel, w = width, h = height, and d = depth. The reference feature maps, $F_{t+1}, F_{t+2}, F_{t+3}, F_{t+4}$, will each be used with the target feature map, F_t , to calculate the attention maps.

The attention map leverages the correlation between the features extracted from a target feature map and a reference feature map. First, we can represent the correlation map as $C = F_t^T \otimes F_{t+i}$, where $\otimes$ represents the matrix multiplication. To utilize the matrix multiplication, F_t and F_{t+i} , where $i = 1, \dots, 4$, are reshaped to the dimension of $c \times whd$. Thus, the correlation map has the dimension of $whd \times whd$.

$$C(F_t, F_{t+i}) = F_t^T \otimes F_{t+i} \in \mathbb{R}^{hwd \times hwd} \quad (1)$$

The correlation map is then normalized with a softmax function to calculate the attention map. Specifically, each entry (j,k) of the attention weight matrix represents the spatiotemporal attention feature between the j location of target feature map, F_t , and the k location of reference feature map, F_{t+i} .

$$A_{t,t+i} = \text{softmax}(C) \in \mathbb{R}^{hwd \times hwd} \quad (2)$$

Once the attention map, $A_{t,t+i}$ is calculated, it is again matrix multiplied by the target feature map, F_t , to ultimately focus in on the regions of the target feature map that show high spatiotemporal correlation with the reference feature map. Ultimately, the post-attention target feature map, $Z_{t,t+i}$ represents refined target feature map given a reference frame in the volume sequence. Figure 2 shows a visual illustration of the attention module.

$$Z_{t,t+i} = F_t \otimes A_{t,t+i} \in \mathbb{R}^{c \times hwd} \quad (3)$$

Once all the post-attention feature maps are calculated, they are concatenated with the original target feature map, F_t , and go forward in a standard decoder phase. As a result, the multi-frame attention maps are trained end-to-end and learn how the reference volume frames can inform the target volume frame to locate the object that is consistently moving through time.

3 Experiments and Results

3.1 Dataset

We evaluate our multi-frame attention approach using in vivo porcine 3D echocardiography dataset. The following volumes were acquired during myocardial infarction studies in pigs. All studies were performed in compliance with the Institutional Animal Care and Use Committee policies. Specifically, the dataset consists of 51 3D+time volumetric echocardiography sequences acquired from 14 pigs. Each pig was imaged at different time points corresponding to different physiological states (baseline rest, 3-day post-balloon occlusion at mid left anterior descending artery (LAD), 7-day post-balloon occlusion). Volumetric sequences were acquired using Philips iE33 ultrasound machine with the X7-2 probe. Each 3D image was taken in the apical view, placing the left ventricle at the center of the image. The endocardium and epicardium boundaries were labeled at the end-systole (ES) and end-diastole (ED) frames by a trained expert. Thus, in total, there are 102 full 3D myocardium labeled in the 51 volumetric sequences. The original image volume size was $224 \times 208 \times 208$. All images were resampled to $64 \times 64 \times 64$ as input to our models.

3.2 Implementation Details

In training our models, we divide our 102 fully labeled in vivo 3D echocardiography frames into 60 training, 20 validation, and 22 testing set. Each sample is either the end-systole (ES) frame or the end-diastole (ED) frame with additional 4 frames that come after the ES or ED frame. The model was trained using a SGD optimizer, batch size of 1, learning rate of 0.0001. A total of 100 epochs were trained. The network is trained by minimizing the binary cross entropy loss function. The model was implemented using PyTorch and the experiments were conducted on a GTX 1080 Ti GPU. Once the model is trained, the inference time is approximately 20s.

3.3 Experimental Studies

We evaluate our multi-frame attention network by comparing with a standard U-Net [15]. To fit the U-Net for our specific 3D echocardiography dataset, we modify the batch normalization into group normalization since our batch size is 1. Group normalization was found to have comparable performance with batch normalization [23], so we believe that the performance of the models would not change significantly when using a higher batch size and batch normalization. We also compare our model with the co-attention network proposed by [10] since it utilizes a similar spatiotemporal attention framework. However, key differences between our proposed method and [10] are (1) our proposed

method utilizes multiple time frames and (2) the output of our proposed method is a single frame segmentation while [10] predicts segmentation masks from both inputs. We also change the number of reference frames to evaluate whether more reference frames would lead to better performance. We conduct our experiments with the same proposed method but with 2 reference frames (3-frame input), 3 reference frames (4-frame input), and 4 reference frames (5-frame input).

To evaluate the accuracy of our proposed method, we adopt the generalized Dice Similarity Coefficient (DSC) and the Volumetric Similarity (VS) [20]. DSC measures the overlap between the predicted mask and the ground truth, while VS considers the how much the volumes of the segments differ. Thus, we selected these metrics that represent different aspects of segmentation accuracy. We also evaluate the correlation coefficient between the predicted ejection fraction and the ground truth ejection fraction using the endocardial boundaries to look at the clinical relevance of our work.

3.4 Results and Discussion

Figure 3 shows a visual illustration of our results. Looking at the U-Net3D prediction and our proposed 5-frame attention model, there is noticeable improvement in the walls where the signal intensity is relatively low compared to other regions (red arrow). This is most likely due to U-Net3D learning the boundaries based on intensity gradient of a single image volume while our proposed method takes into account the movement of the myocardial boundaries to guide the location of the walls even when the signal intensity is low. Table 1 reports the complete findings of our results. Our findings show that utilizing multiple time frames in a given sequence to segment the left ventricle at our target frame improves the overall performance. It is likely that the high temporal resolution of echocardiography gives the model better context as to where the walls are at any one time point by utilizing other frames. Moreover, increasing the number of reference frames to capture additional spatiotemporal attention features augments the performance even further. This is also seen in Fig. 3, where in the short axis view images (first row) among 3 Frame, 4 Frame, 5 Frame, the 5 Frame model completes the doughnut shape that matches well with the ground truth. Table 2 reports the analysis of statistical significance between the evaluated methods based on a paired t-test. It shows that there is a statistically significant improvement ($p < 0.05$) in the Dice Similarity Coefficient of multi-frame (5 frames) method when compared to multi-frame (4 frames) method. Since multi-frame (4 frames) outperforms U-Net3D and Co-Attention methods, we can infer that multi-frame (5 frames) performs the best based on the p -value. The Volumetric Similarity (VS) values did not show statistically significant improvement when comparing the different number of reference frames in the multi-frame method. However, when comparing the performance between previous methods (U-Net3D, Co-Attention), multi-frame method showed statistically significant difference.

To analyze the clinical relevance and use of our proposed method, we also look at the correlation coefficient between the ground truth ejection fraction (EF) and the predicted ejection fraction. Our results show that the EF correlation coefficient increases with the use of multi-frame attention features.

3.5 Limitations and Future Works

Although the proposed multi-frame attention network shows promise in improving segmentation tasks in a spatiotemporal dataset such as 3D echocardiography, it has much room for improvement. First, the current study evaluated up to 5 time frames as maximum into the model. We will explore incorporating more time frames, potentially including all time frames in a given echocardiography sequence. Second, our dataset is currently limited to the porcine in vivo models that were acquired during a myocardial infarction studies. These images have a pigtail catheter inserted in the left ventricle which may introduce artifacts in the image. Also, some images were acquired during open chest thoracotomy procedure, while other images were acquired during a closed chest state. Therefore, in future studies, we will incorporate 3D echocardiography sequences from patients to extend our framework in a fully standardized clinical setting.

4 Conclusion

In this paper, we present a multi-frame attention architecture for segmentation of left ventricle in 3D echocardiography. The multi-frame attention network successfully utilized the spatiotemporal features of echocardiography image volumes in a given sequence, and it extracted highly correlated features between target volume and reference volume to guide the location of the myocardial boundaries even when the signal intensity was low. Our proposed framework is also easily scalable to incorporate as many frames as reference volumes. The results indicate that taking advantage of the spatiotemporal nature of images can augment the performance of 3D left ventricle segmentation. Future work will expand our validations to include clinical 3D echocardiography datasets to train our models more robustly. Also, we will evaluate the effectiveness of attention using public 2D echocardiography datasets to see the feasibility of multi-frame attention network for different data dimensions.

Acknowledgments.

We would like to thank the technical assistance provided by the staff of the Yale Translational Research Imaging Center. This work was supported in part by the following grants: R01HL121226, R01HL137365, F30HL158154, and Medical Scientist Training Program Grant T32GM007205.

References

1. Ahn SS, Ta K, Lu A, Stendahl JC, Sinusas AJ, Duncan JS: Unsupervised motion tracking of left ventricle in echocardiography. In: Medical Imaging 2020: Ultrasonic Imaging and Tomography, vol. 11319, p. 113190Z. International Society for Optics and Photonics (2020)
2. Chen C., et al. : Deep learning for cardiac image segmentation: a review. Front. Cardiovasc. Med 7, 25 (2020) [PubMed: 32195270]
3. Compas CB, et al. : Radial basis functions for combining shape and speckle tracking in 4d echocardiography. IEEE Trans. Med. Imaging 33(6), 1275–1289 (2014) [PubMed: 24893257]
4. Dong S., et al. : Deep atlas network for efficient 3d left ventricle segmentation on echocardiography. Med. Image Anal 61, 101638 (2020) [PubMed: 32007701]
5. Dong S, Luo G, Wang K, Cao S, Li Q, Zhang H.: A combined fully convolutional networks and deformable model for automatic left ventricle segmentation based on 3d echocardiography. BioMed Res. Int 2018 (2018)

6. Ghorbani A., et al. : Deep learning interpretation of echocardiograms. *NPJ Dig. Med* 3(1), 1–10 (2020)
7. Huang X., et al. : Contour tracking in echocardiographic sequences via sparse representation and dictionary learning. *Med. Image Anal* 18(2), 253–271 (2014) [PubMed: 24292554]
8. Leclerc S., et al. : Deep learning for segmentation using an open large-scale data set in 2d echocardiography. *IEEE Trans. Med. Imaging* 38(9), 2198–2210 (2019) [PubMed: 30802851]
9. Liu F, Wang K, Liu D, Yang X, Tian J.: Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography. *Med. Image Anal* 67, 10187 (2021)
10. Lu X, et al.: See more, know more: unsupervised video object segmentation with co-attention siamese networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3623–3632 (2019)
11. Luong MT, et al. : Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015)
12. Ouyang D., et al. : Video-based ai for beat-to-beat assessment of cardiac function. *Nature* 580(7802), 252–256 (2020) [PubMed: 32269341]
13. Papademetris X, Sinusas AJ, Dione DP, Constable RT, Duncan JS: Estimation of 3-d left ventricular deformation from medical images using biomechanical models. *IEEE Trans. Med. Imaging* 21(7), 786–800 (2002) [PubMed: 12374316]
14. Parajuli N., et al. : Flow network tracking for spatiotemporal and periodic pointmatching: Applied to cardiac motion analysis. *Med. Image Anal* 55, 116–135 (2019) [PubMed: 31055125]
15. Ronneberger O, Fischer P, Brox T: U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF (eds.) *MICCAI 2015*. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015).
16. Schlemper J., et al. : Attention gated networks: learning to leverage salient regions in medical images. *Med. Image Anal* 53, 197–207 (2019) [PubMed: 30802813]
17. Shi P, Sinusas AJ, Constable RT, Ritman E, Duncan JS: Point-tracked quantitative analysis of left ventricular surface motion from 3-d image sequences. *IEEE Trans. Med. Imaging* 19(1), 36–50 (2000) [PubMed: 10782617]
18. Stendahl JC, et al. : Regional myocardial strain analysis via 2d speckle tracking echocardiography: validation with sonomicrometry and correlation with regional blood flow in the presence of graded coronary stenoses and dobutamine stress. *Cardiovasc. Ultrasound* 18(1), 1–16 (2020) [PubMed: 31931808]
19. Ta K, Ahn SS, Lu A, Stendahl JC, Sinusas AJ, Duncan JS: A semi-supervised joint learning approach to left ventricular segmentation and motion tracking in echocardiography. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1734–1737. IEEE (2020)
20. Taha AA, Hanbury A.: Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Med. Imaging* 15(1), 1–28 (2015) [PubMed: 25645550]
21. Virani SS, et al. : Heart disease and stroke statistics-2020 update: a report from the American heart association. *Circulation* 141, E139–E596 (2020) [PubMed: 31992061]
22. Wu L., et al. : Deep coattention-based comparator for relative representation learning in person re-identification. *IEEE Trans. Neural Netw. Learn. Syst* 32, 722–735 (2020)
23. Wu Y, He K: Group normalization. In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19 (2018)
24. Xu H, Saenko K: Ask, attend and answer: exploring question-guided spatial attention for visual question answering. In: Leibe B, Matas J, Sebe N, Welling M (eds.) *ECCV 2016*. LNCS, vol. 9911, pp. 451–466. Springer, Cham (2016).
25. Xu K, et al.: Show, attend and tell: neural image caption generation with visual attention. In: *International Conference on Machine Learning*, pp. 2048–2057 (2015)
26. Zhang J., et al. : Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation* 138(16), 1623–1635 (2018) [PubMed: 30354459]

Fig. 1.
Proposed Multi-frame Attention Network for segmentation of 3D left ventricle in echocardiography.

Fig. 2.
Visual illustration of Multi-frame Attention Module.

Fig. 3.

Cross sectional display of echocardiography images and left ventricle masks (blue) at 3 different viewing planes. The results shown in this figure are from the same volumetric image. Red arrow shows that incorporating multiple time frames as reference frames allows the model to identify regions that are difficult to pick up from intensity information alone.

(Color figure online)

Table 1.

Summary of segmentation evaluation metrics. DSC = Dice Score Coefficient, VS = Volumetric Similarity, EF Corr Coeff = Ejection Fraction Correlation Coefficient. Higher number is better for all metrics. Myocardium refers to the combined endocardial and epicardial boundaries of the left ventricle.

Methods	DSC (Myocardium)	VS (Myocardium)	EF correlation coeff.
U-Net3D	0.451 (± 0.08)	0.61 (± 0.15)	0.231
Co-Attention ([10])	0.545 (± 0.06)	0.43 (± 0.11)	0.257
Multi-frame (3 frames)	0.582 (± 0.14)	0.88 (± 0.08)	0.305
Multi-frame (4 frames)	0.572 (± 0.16)	0.86 (± 0.14)	0.486
Multi-frame (5 frames)	0.641 (± 0.14)	0.90 (± 0.08)	0.673

Table 2.

Summary of paired t-test for statistical significance between methods.

Comparison	DSC	VS
U-Net3D vs. CoAttention [10]	p = 0.02	p = 0.001
Co-Attention ([10]) vs. Multi-frame (3 frames)	p = 0.08	p = 5.27e-10
Multi-frame (3 frames) vs. Multi-frame (4 frames)	p = 0.89	p = 0.37
Multi-frame (4 frames) vs. Multi-frame (5 frames)	p = 0.01	p = 0.25