Project Milestone- Echocardiography Segmentation and Ejection Fraction Estimation

Cheng Che Tsai
Department of Computer Science, University of North Carolina at Chapel Hill
cctsai@cs.unc.edu

Alessandro Folloni
Division of Artificial Intelligence, Alma Mater Studiorum Università Di Bologna
afolloni@unc.edu

Abstract

Our project focuses on providing fast and comprehensive solutions for echocardiography analysis. Specifically, we aim to build a model for left ventricle segmentation and ejection fraction estimation based on video analysis techniques learned in class. Past studies have thoroughly explored these two topics, but most of them were based on image analysis techniques instead of video's. These image-based techniques limit most of these tasks to take the key frames (the end frames for systolic and diastolic phases) as input, which require pre-selection by human experts and lose abundant information contained in the frames in-between. In this project, we aim to build a video-based model that utilizes entire uncurated videos for simultaneously providing video segmentation for left ventricle and estimation for ejection fraction.

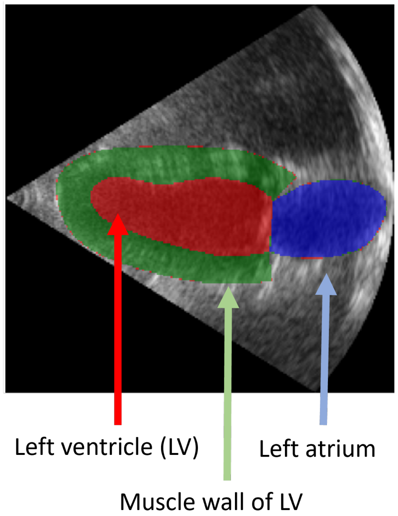

Figure 1. Segmentation targets of the heart

1. Problem Definition

1.1. Echocardiography

Echocardiography, often referred to as ”heart echo,” is a valuable medical imaging technique for assessing heart structure and function. It aids in early heart problem detection, monitors cardiac conditions, and assesses the effectiveness of medical interventions. Compared to heart CT and MRI, echocardiography is fast and radiation-free. However, echocardiography has two major limitations: the diagnostic value is operator-dependent, and interpreting a heart echo takes a long work time from the cardio experts.

These two limitations underscore the need for a standardized, non-operator-dependent method to provide automatic heart echo analysis, which is the motivation behind our project. Specifically, we aim to focus on two key indices of the standard heart evaluation: left ventricle (the major chamber of the heart) segmentation and ejection fraction estimation (a regression task).

1.2. Left Ventricle Segmentation (LV Segmentation)

Left ventricle is the main chamber of our heart Fig 1. LV segmentation provides valuable information for several disease diagnosis. In fact, it is equally important to have masks for both muscular wall of LV and for left atrium. We'll explore all these segmentation in the final report. For the milestone, we'll only provide our result for LV segmentation.

1.3. Ejection Fraction Estimation (EF estimation)

Ejection fraction estimates how well your heart pumps blood. It is defined as the percentage of blood pumped out of the heart divided by its initial volume per heart stroke. A healthy heart typically has an EF of around 55% to 70%

and values below 50% will be considered as heart failure. Mathematically, it can be defined as:

$$EF = \frac{EDV - ESV}{EDV} \times 100\%$$

where

- EDV stands for End-Diastolic Volume, which is the amount of blood in the left ventricle at the end of diastole (when the heart is fully relaxed and filled with blood).

- ESV stands for End-Systolic Volume, which is the amount of blood in the left ventricle at the end of systole (when the heart has just contracted and pumped out blood).

1.4. Overall Goals

Obviously, LV segmentation with temporal coherence, when used together, provides a way to estimate the volume size of the LV and is directly related to EF estimation. We believe these two tasks can benefit from each other if combined in a single model.

Therefore, our goal is to build a model that can simultaneously provide the segmentation mask for the left ventricle and estimate ejection fraction (and possibly EDV and ESV) in one step. The input to the model will be short clips of heart echo. These video clips can contain either a single cycle of heart beating or multiple cycles. We'll build and test our models on both clip configurations.

2. Related Work

2D CNNs: In 2019, Leclerc et al. [1] established a milestone in echocardiographic segmentation by publishing the CAMUS dataset. This is the first publicly available larger-scale dataset for performing segmentation on heart echo. The authors also demonstrated competitive results with different UNet-based models. A year later, Leclerc et al. [2] introduced their LU-Net, a two-step segmentation network inspired by Mask R-CNN. LU-Net showed improved results compared to the authors' previous UNet architecture, even beating intra-observer accuracy on the epicardium.

2D+time CNNs: Alongside the work of Leclerc et al., Ouyang et al. introduced their EchoNet-Dynamic dataset [3]. Initially developed as a single-task model using the R(2+1)D architecture, they later enhanced its performance by transforming it into a multi-task model for LV segmentation and EF regression [4], while constraining it to a single beat cycle. The EF regression achieved an MAE of 4.10.

More recently, Wei et al. introduced CLAS [5], a unique 3D segmentation network aimed at achieving temporal consistency with only ED and ES annotations. To achieve this, CLAS predicts deformation fields and utilizes them for annotation propagation during training, improving the consistency between ED and ES predictions. However, their evaluation of temporal consistency across sequences was primarily qualitative, focusing on a limited number of patients. Dai et al. [6] proposed AdaCon, a novel contrastive learning framework for regression problems. These end-to-end regression methods yield more reliable estimates, avoiding error propagation from separate frames, resulting in a 3.3% reduction in MAE.

Limitations of past approaches: There are three major limitations in the past studies that we aim to address and improve:

- First, most EF estimation models discard the intermediate frames between end-diastole (ED) and end-systole (ES), even though these frames contain valuable information for characterizing other pathologies and can synergize with LV segmentation.

- Second, the models used to model temporal information have all been CNN-based models. However, these models are not considered state-of-the-art in the era of transformers.

- Third, these models have not fully leveraged the repetitive nature of the data. Video clips between cardiac cycles represent a natural augmentation that can be harnessed to enhance performance. An illustrative example can be found in [7].

Hence, we aim to develop a model that can provide LV segmentation masks across the videos, estimate the dynamic volumes of the beating heart, and provide EF estimation in the end.

3. Technical Approach

3.1. Architectures overview

In the proposal, we planned to explore two popular methods for building our model: TimeSformer[8] and R-CNN[9]. TimeSformer is known for its excellent ability in formatting temporal attention, and R-CNN is a very standard model to consider when dealing with semantic segmentation. However, the publicly available TimeSformer on GitHub is primarily trained for video classification. In other words, it lacks the decoder part for reconstructing segmentation masks from the compressed token in the feature spaces. Unfortunately, we failed to train our version of the decoder for the pre-trained TimeSformer. Additionally, we decided to first test our notion of implementing temporal motion coherence and segmentation-based EF estimation. Therefore, instead of modifying R-CNN as an alternative, we decided to start with the simplest model we could get: the DeepLabv3 [10].

DeepLabv3 was a then-state-of-the-art semantic segmentation model, excelling in capturing fine-grained details and intricate object boundaries within images. A key innovation of DeepLabv3 lies in its employment of dilated convolution, enabling the model to maintain a large receptive field while preserving spatial resolution. Furthermore, DeepLabv3 incorporates the Spatial Pyramid Pooling module, facilitating multi-scale feature representation and accommodating diverse object sizes within an image.

As the feedback in the class presentation said, DeepLabv3 is outdated in 2023. We searched the paper with code website, and list out some of the models that interest us for replacing DeepLabv3 in the future.

- TMANet [11]
- TDNet [12]
- Tube-link [13]

3.2. Our designs

Building upon the foundation of DeepLabv3, we introduced a new input to enable the model to perceive time in LV segmentation. Additionally, we incorporated two specific designs tailored for more accurate EF estimation.

Temporal geometric constraint for semantic targets: The heart is beating smoothly. That means ensuring smoothness in the movement of the heart across consecutive frames is crucial for accurate and clinically meaningful results. To address this, we introduce a temporal smoothness constraint to our video model. A Temporal Smoothness Loss (TS loss) penalizes deviations in intensity between adjacent frames:

$$\text{TS loss} = \frac{1}{N-1}\sum_{i=1}^{N-1} \|\text{frames}[i+1] - \text{frames}[i]\|^2$$

Please note that, different from conventional smoothness loss on images, we impose the losses on each semantic mask separately. We intend to ensure the smoothness of each organ, not the general image.

Mask-based volume estimation: Previous models directly estimate the ED and ES volumes by their corresponding single-frame image. This approach requires manual annotation for the locations of ED and ES and do not consider the temporal information (blood flow, physiologically) between the frames. Contrast to the previous approaches, our model estimate EF in the following steps:

- Taking segmentation masks of LV as inputs, our model provide LV volume estimate for each frame. We also us L2 loss to penalize drastic changes in volume estimation for adjacent frames.
- The model then use max and min pooling for knowing the maximum and minimal volumes, and therefore uses them to compute the EF.

Please also refer to the 'future plan' section below to see the advantages and reasons why we'd use semantic masks for EF estimation.

4. Experiments

4.1. Evaluation

LV segmentation:we will evaluate our segmentation results on three metrics.

- The first one is the commonly used Dice score evaluated only at the end of the systolic phase and at the end of the diastolic phase, which is the conventional methods used in the previous study.
- The second one is the temporal IoU [14]. This helps to establish the temporal evaluation of the generated masks.
- The third metric is a second derivative-based image smoothness score. This score can help us to understand if the temporal segmentation masks is anatomically reasonable.

EF regression: We will evaluate model performance using the most commonly used metrics: R^2, MAE and RMSE.

Expectation: Regarding our expectations, we do not expect our model can perform better at the systolic/diastolic phase evaluation since past models have provided a decent 0.95 performance. However, we expect our model provide better smoothness between the frames and better temporal IoU because of incorporating the temporal information and the beat-cycle constraints imposed.

Regarding EF regression, past studies heavily relied on knowing the positions of the end-diastole (ED) and end-systole (ES) phases. Without this information, the R^2 value dropped significantly. Therefore, we anticipate that our model can substantially advance the boundaries of EF estimation.

4.2. Dataset

For our project, the choice of dataset for use is not stragithforward as the first glance. There are not many validated and meaningful sources in the literature, indeed. We opted for two datasets that were relatively popular in the past and suit perfectly for our goals: CAMUS and EchoNet-Dynamic.

CAMUS dataset (Cardiac Acquisitions for Multi-structure Ultrasound Segmentation) was introduced in 2019. It contains 2D echocardiographic sequences with two and four-chamber views (two different angles of viewing

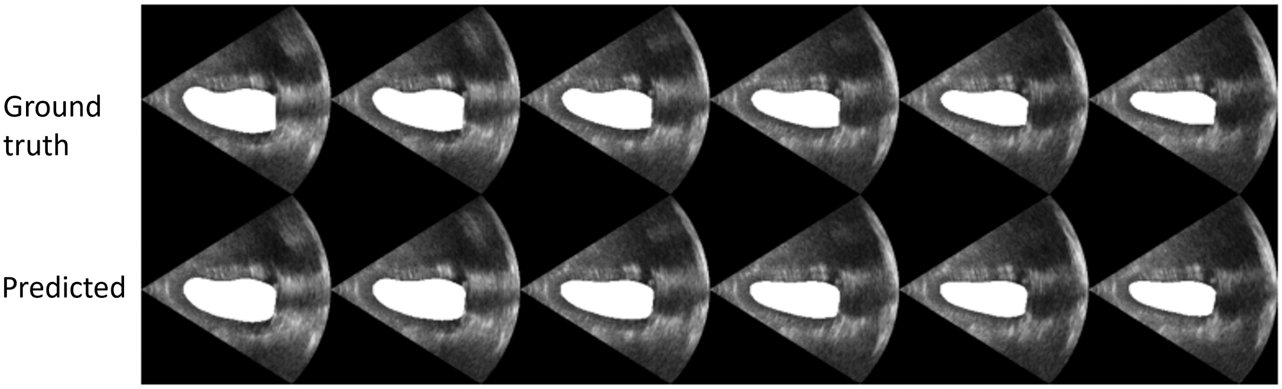

Figure 2. Our result

the heart) of 500 patients acquired at the University Hospital of Saint Etienne. Three cardiologists annually annotated the left ventricle endocardium, the myocardium, and the left atrium. In other words, we have keypoints of the inner and outer walls of the left ventricle.

Moreover, it has the peculiarity of having one subset annotated by the same physician 7 months apart; this is helpful to measure intra- and inter-operator variability. Regarding the dimensionality, there is not a fixed frame size; each video has a different resolution (all of them are larger than 1024 × 512). This is due to the fact that it enforced clinical realism. That is to say, there was no pre-processing on the data acquired when creating the dataset.

EchoNet-Dynamic was created in 2021 to provide images for comprehensive study of cardiac motion and chamber volumes using echocardiography on a large scale. These videos were acquired during real clinical practice for diagnostic and medical decision-making purposes. The dataset comprises 10,036 videos, each originating from a different individual. A typical full resting echocardiogram study consists of a series of 50-100 videos and still images, capturing various perspectives of the heart through different angles, locations, and image acquisition techniques. Ultimately, from each study, one apical-4-chamber 2D grayscale video is extracted. The uniqueness of the dataset is that each video is linked to clinical measurements and calculations and it is full of different annotations including paired electrocardiography (ECG) and ejection fraction.

These videos depict canonical apical-4-chamber views or zoomed-in apical-4-chamber echocardiographic views of sufficient quality for sonographers to determine left ventricular volumes as part of their standard clinical workflow. Each video underwent cropping and masking to remove medical text and information outside the scanning sector. Subsequently, all resulting images were downsized to a standard 112x112 size.

Method	temporal IoU
Vanilla	0.753
Ours	0.81

Table 1. Comparison between the vanilla DeepLabv3 and our proposed model shows the effectiveness of temporal smoothness loss.

5. Result

Using the CAMUS dataset, we trained the vanilla DeepLabv3 model along with our proposed model to see the effectiveness of the proposed temporal smoothness loss and EF estimation method.

5.1. LV segmentation

In Table 1, the our proposed model provides a 5% more temporal IoU compared to the vanilla DeepLabv3, which shows the effectiveness of temporal smoothness loss in training. Please note that the segmentation measurement obtained here seems much lower than previous relevant works at first glance. However, previous reported Dice scores were computed only on ED and ES and the locations of ED and ES were given for training their models. Our model instead only takes the video as input and the temporal masks as the supervised signals. A fair comparison should be conducted between the vanilla DeepLabv3 and our proposal as shown in Table 1.

5.2. EF estimation

Unlike the success achieved in LV segmentation, our proposed model resulted in a higher RMSE of 17.4, significantly surpassing the previous best of 5.17. This disparity is notable, considering that previous researchers employed a direct approach using ED and ES frames to estimate EF, a strategy not permissible in our model. This discrepancy suggests potential issues with our approach.

Primarily, the utilization of max-pooling and min-pooling may not be optimal. While these operations demonstrate robustness to video inputs cropped from any point in the heart cycle, they hinder the backpropagation of gradients to frames providing intermediate values in-between. To address this, alternative methods should be explored, such as directly using all masks as feature maps to estimate a scalar value as output. Additionally, combining multiple semantic masks to estimate EF is another avenue worth investigating. Further details can be found in the 'future plan' section below.

5.3. Future plans

DeepLabv3 is an outdated model in 2023. Our first goal is to modify a state-of-the-art video model (TMANet, TDNet, or Tube-link) for our task.

Unlike the success observed in LV segmentation, our model's performance in EF estimation is notably inadequate. To address this limitation, an additional direction worth exploring involves incorporating more semantic masks for EF estimation. The underlying concept is rooted in the conservation of blood flow. Throughout a heart cycle, when the left atrium pumps blood into the left ventricle, the total volume of the LA and LV should remain relatively constant. Therefore, we hypothesize that introducing the segmentation masks of the left atrium as input could potentially enhance both LV segmentation and EF estimation.

The implementation of the blood volume conservation law is not as straightforward as it may appear. This conservation principle is applicable only during specific periods within the heart cycle, rather than throughout the entire cycle. To accurately enforce this volume conservation law, we require annotations indicating the heart phase for each video frame. Unfortunately, such annotations are not available in the provided data. Thus, we must devise alternative methods that can furnish information about the heart cycle for each frame to successfully integrate this design into our model.

References

[1] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, *et al.*, "Deep learning for segmentation using an open large-scale dataset in 2d echocardiography," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2198–2210, 2019. 2

[2] S. Leclerc, E. Smistad, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, M. Belhamissi, S. Israilov, T. Grenier, *et al.*, "Lu-net: a multistage attention network to improve the robustness of segmentation of left ventricular structures in 2-d echocardiography," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, pp. 2519–2530, 2020. 2

[3] D. Ouyang, B. He, A. Ghorbani, M. P. Lungren, E. A. Ashley, D. H. Liang, and J. Y. Zou, "Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning," in *NeurIPS ML4H Workshop: Vancouver, BC, Canada*, 2019. 2

[4] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley, *et al.*, "Video-based ai for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, pp. 252–256, 2020. 2

[5] H. Wei, H. Cao, Y. Cao, Y. Zhou, W. Xue, D. Ni, and S. Li, "Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, pp. 623–632, Springer, 2020. 2

[6] W. Dai, X. Li, W. H. K. Chiu, M. D. Kuo, and K.-T. Cheng, "Adaptive contrast for image regression in computer-aided disease assessment," *IEEE Transactions on Medical Imaging*, vol. 41, no. 5, pp. 1255–1268, 2021. 2

[7] W. Dai, X. Li, X. Ding, and K.-T. Cheng, "Cyclical self-supervision for semi-supervised ejection fraction prediction from echocardiogram videos," *IEEE Transactions on Medical Imaging*, 2022. 2

[8] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *ICML*, vol. 2, p. 4, 2021. 2

[9] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015. 2

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017. 2

[11] H. Wang, W. Wang, and J. Liu, "Temporal memory attention for video semantic segmentation," in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 2254–2258, IEEE, 2021. 3

[12] X. Xu, G. Geng, X. Cao, K. Li, and M. Zhou, "Tdnet: transformer-based network for point cloud denoising," *Applied Optics*, vol. 61, no. 6, pp. C80–C88, 2022. 3

[13] X. Li, H. Yuan, W. Zhang, G. Cheng, J. Pang, and C. C. Loy, "Tube-link: A flexible cross tube baseline for universal video segmentation," *arXiv preprint arXiv:2303.12782*, 2023. 3

[14] L. Yang, Y. Fan, and N. Xu, "Video instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5188–5197, 2019. 3