Assessing the Generalizability of Temporally-Coherent Echocardiography Video Segmentation

Yida Chen[a], Xiaoyan Zhang[b], Christopher M. Haggerty[b], and Joshua V. Stough[a]

[a]Computer Science, Bucknell University, Lewisburg, PA;
[b]Translational Data Science and Informatics, Geisinger, Danville, PA

ABSTRACT

Existing deep-learning methods achieve state-of-art segmentation of multiple heart substructures from 2D echocardiography videos, an important step in the diagnosis and management of cardiovascular disease. However, these methods generally perform frame-level segmentation, ignoring the temporal coherence in heart motion between frames, which is a useful signal in clinical protocols. In this work, we implement temporally consistent video segmentation, which has recently been shown to improve performance on the multi-structure annotated CAMUS dataset. We show that data augmentation further improves results, which are consistent with prior state-of-art works. Our 10-fold cross-validation shows that video segmentation improves the automatic comparison to clinical indices including smaller median absolute errors for left ventricular end-diastolic volume (6.4 ml), end-systolic volume (4.2 ml), and ejection fraction (EF) (3.5%). In segmenting key cardiac structures, video segmentation achieves mean Dice overlap of 0.93 on left ventricular endocardium, 0.95 on left ventricular epicardium, and 0.88 on left atrium. To assess clinical generalizability, we further apply the CAMUS-trained video segmentation models, without tuning, to a larger, recently published EchoNet-Dynamic clinical dataset. On 1274 patients in the test set, we obtain a median absolute error of $4.9\% \pm 5.4$ in EF, confirming the reliability of this scheme. In that the EchoNet-Dynamic videos contain limited annotation only for left ventricle endocardium, this effort extends at little cost generalizable, multi-structure video segmentation to a large clinical dataset.

Keywords: Echocardiography, Segmentation, Quantitative Image Analysis, Neural Networks

1. INTRODUCTION

Accurate quantification of clinical indices of cardiovascular diseases (CVDs) requires precise annotations of the key cardiac structures, including the left ventricle endocardium (LV_{endo}), epicardium (LV_{epi}), and left atrium (LA). Manual annotation requires clinicians to locate the end-diastolic (ED) and end-systolic (ES) frames in an echocardiographic video and carefully delineate the cardiac structures in the noisy images by visual inspection. The recommended method of measuring the EF from echocardiogram relies on the manually annotated ED and ES frames in both apical two chamber (AP2) and four chamber (AP4) views. Consequently, measurements of EF in clinic are time-consuming and vulnerable to a high inter-observer variability among human experts,[1] motivating the development of automatic techniques.[2]

Existing deep learning methods achieve accurate segmentation through independently analyzing a video's constituent frames and determining the ED and ES phases.[2–4] However, clinicians often inspect the whole echo videos to identify the possible ED/ES frames, and use temporal coherence between the video frames to assist the manual annotations. Recent works in magnetic resonance[5,6] and echocardiography[4,7,8] use whole or cropped videos as the inference source and more closely simulate the clinical annotation process. Qin et al.[6] and Li et al.[7] use recurrent units to model the temporal relationship between segmented frames. Ouyang et al.[4] use frame-level segmentation to determine ED-ES video clips, then regress on EF using spatio-temporal (spatial 2D + temporal 1D) convolutions. In their work, they also publish the large EchoNet-Dynamic dataset of multi-beat videos with limited LV_{endo} annotations in AP4 view.

Corresponding author: yc015@bucknell.edu, joshua.stough@bucknell.edu

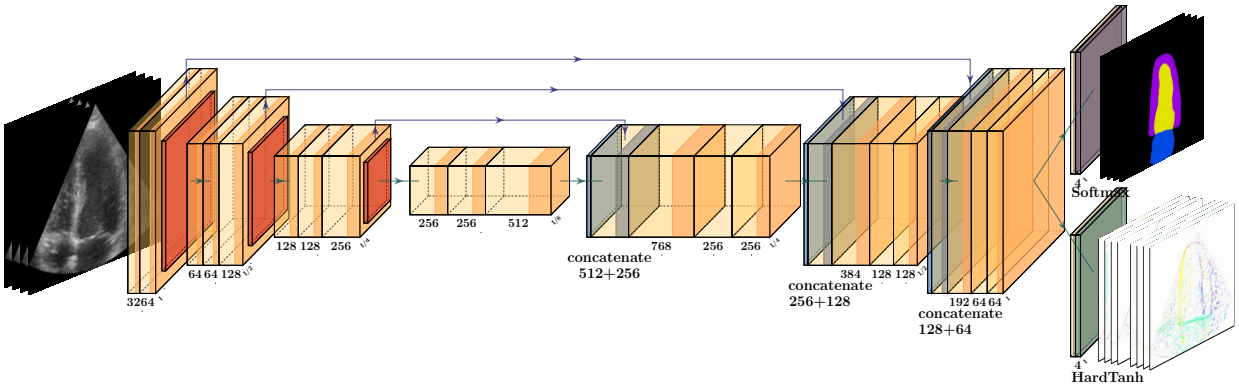

Figure 1. CLAS[8] joint motion tracking and video segmentation. The 3D U-Net serves as a shared features extractor. We normalize the motion tracking output with HardTanh and segmentation output with Softmax.

Recent work from Wei et al.[8] uses a 3D U-Net[9] as the shared feature extractor for dual motion tracking and video segmentation tasks, called CLAS. They report excellent results on the public CAMUS dataset,[10] which contains annotations for three substructures (LV_{endo}, LV_{epi}, LA) and single ED-ES video clips. In this work we compare CLAS to a prior state-of-the-art frame-level segmentation method[3] using the same 10-fold cross validation study framework with data augmentation. We use Dice overlap and derived LV_{endo} volumes and EF to quantify the difference in performance. We further assess the generalizability of CLAS[8] trained on CAMUS to the EchoNet-Dynamic dataset, achieving results consistent with those of Ouyang et al.[4] but with multi-structure segmentation of ED-ES video clips within each video.

2. ARCHITECTURE AND METHODS

In our implementation of the CLAS[8] video segmentation network(Fig. 1), we used the 3D U-Net proposed by Çiçek et al.[9] as the shared feature extractor for the motion tracking and video segmentation tasks. The 3D U-Net takes in the ordered sequence of echocardiographic frames from ED to ES phases (ED-ES video clip, or clip) and outputs a 64 channel features map with the same height, width, and temporal depth as the input clip. This feature map is then sent to a segmentation head with a single $1 \times 1 \times 1$ convolutional layer and softmax to acquire the 4-channel frame-level segmentation of the clip, supervised with a combined cross-entropy and multi-class Dice loss. The same feature map is also processed by a motion tracking head with a single $3 \times 3 \times 3$ convolutional layer (padding=1) to obtain the bi-directional motion fields (forward and backward motions) between the frames in the input clip. The motion field outputs are supervised by a combination of local cross correlation loss[11] (number of local windows n=16) and smoothness loss.[12] The weights are initially trained using the above segmentation and motion losses in a warm start.

In the CAMUS dataset, the clips are sparsely annotated such that only the ED and ES frames have the corresponding ground true labels. The bi-directional motion fields allow for the spatial transformation of the one-hot encoded ground truth ED or ES labels from the ends of each clip, generating pseudo labels for intermediate frames. The motion tracking and segmentation heads are then tied together through additional multi-class Dice losses: first, between corresponding intermediate frames from both the tracking and segmentation heads; and second, between the motion-transformed ED (forward) and ES (backward) labels and the ground truth ES and ED labels respectively. Tying together these additional pseudo label transformations reinforces coherency between the segmentation and motion tracking outputs.

2.1 Data Augmentation on Echocardiogram Videos

Stough et al.[3] showed the effectiveness of the data augmentation in enhancing the performance of frame-level segmentation in the CAMUS set. We perform a similar set of data augmentations, including the random intensity windowing, slight rotation about the transducer point, and additive Gaussian noise, applied on the fly to the clips during the training. The subintervals of the intensity windowing and the angles of rotation are variant

across the patients but identical within a clip. The scale of the added Gussian-distributed noise is random on each frame of a clip.

2.2 Training Setup

Ten frames from ED to ES were resampled with equal time interval from each echocardiogram video. All frames were resized to 256×256 and the intensities normalized to [-1, 1]. We trained the separate models on the AP2 and AP4 views for 40 epochs with Adam optimizer, saving the model weights with the best loss on the validation set. The initial learning rates of the Adam optimizer on the 3D U-Net feature extraction and segmentation head were 1×10^{-4}, and 0.5×10^{-4} on the motion tracking head. The learning rate of all optmizers were reduced to 1×10^{-5} after 25 epochs. The weights of all 3D convolutional layers were initialized according the Gaussian $\mathcal{N}(0, 10^{-5})$. For data augmentation, the subinterval of the random intensity windowing was up to half of the image's intensity range; the random rotation about the transducor was sampled from $\mathcal{N}(0, \sigma_{rot}^2 = 5^2)$; and the scale of additive Gaussian noises was from sampled from $\mathcal{U}(0, 0.15)$. We used the same set of hyperparameters across the training folds.

2.3 Evaluation

We used Dice overlaps to validate the methods' performance on the CAMUS dataset. Dice overlap measures the agreement between the automated and manual annotations by the ratio of the intersection to average area, $D(y_{auto}, y_{true}) = \frac{2(y_{auto} \cap y_{true})}{|y_{auto}| + |y_{true}|}$.

Since both the AP2 and AP4 views of echocardiograms are available in the CAMUS dataset, we used the Simpson's biplane method of disk, the most accurate protocol,[13] to derive the ED and ES volumes (EDV and ESV) from the automated segmentations. This protocol approximates the left ventricle as contiguous elliptical cylinders of which the semi-major and semi-minor axes are estimated by the widths of left-ventricle in AP2 and AP4 views respectively. The derived EF of a patient is equal to $\frac{\text{EDV}-\text{ESV}}{\text{EDV}} \times 100\%$. With only AP4 views available in the EchoNet-Dynamic dataset however, Simpson's monoplane assumes contiguous circular cylinder disks in the geometric approximation.

3. EXPERIMENTAL RESULTS

The CAMUS dataset consists of 450 echocardiogram patients in the training set with videos in both AP2 and AP4 views and manual annotated frames in ED and ES phases (900 echocadiogram videos). Each video contains one half heartbeat clip from ED to ES phase with length of at least 10 frames. We performed a ten-fold cross-validation experiment with folds split by patient, so that each patient is represented in *exactly one fold*. Sampling is further stratified on EF range ($\leq 45\%, \geq 55\%, else$) and reported AP2 image quality (good, medium, poor), as previously suggested.[3,10]

Table 1 shows Dice performance by phase and heart substructure for three methods. Our implementation of CLAS is consistent with Wei et al.,[8] and data augmentation (A-CLAS) provides slight but significant further improvement. A-CLAS does show slightly lower performance than the U-Net based frame-level segmentation of Stough et al.[3] at the ED and ES frames, for which annotations are available. However, A-CLAS forces a temporally coherent segmentation tying together the ED and ES phases, which may negatively impact Dice for

Methods	Dice - ED			Dice - ES		
	$\text{LV}_{endo} \pm \sigma$	$\text{LV}_{epi} \pm \sigma$	$\text{LA} \pm \sigma$	$\text{LV}_{endo} \pm \sigma$	$\text{LV}_{epi} \pm \sigma$	$\text{LA} \pm \sigma$
CLAS[8]	0.935 ± 0.034	0.948 ± 0.027	0.857 ± 0.118	0.911 ± 0.052	0.941 ± 0.031	0.892 ± 0.086
A-CLAS (w/ aug)	0.937 ± 0.034	0.950 ± 0.028	0.863 ± 0.112	0.916 ± 0.050	0.943 ± 0.031	0.898 ± 0.082
Frame[3] (w/ aug)	0.941 ± 0.030	0.956 ± 0.022	0.872 ± 0.120	0.916 ± 0.057	0.948 ± 0.040	0.909 ± 0.084

Table 1. Mean Dice overlaps ($\pm$ standard deviation) on the 450 training patients in CAMUS using 10-fold cross-validation. CLAS[8] is shown both without and with data augmentation. The paired Wilcoxon signed-rank test[14] on corresponding frames indicates that data augmentation improves CLAS performance on the three cardiac structures ($p \ll 0.001$).

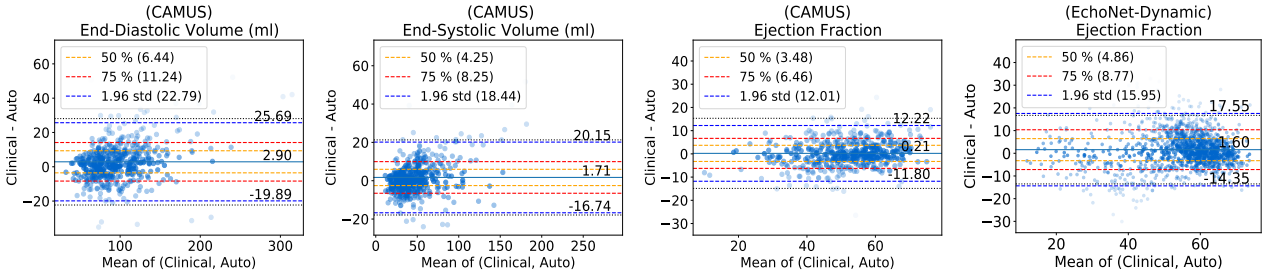

Figure 2. Bland-Altman plots of derived clinical indices on the CAMUS dataset, and EF on EchoNet-Dynamic. The [orange, red] dashed lines show bias ± [50%,75%]. The blue dotted line shows bias $\pm 1.96\sigma$. The black dotted line is the inter-observer variability in non contrast echocardiography (-7.6 ± 25.2 EDV; -3.5 ± 19.6 ESV; -1.2 ± 15.1 EF) .[1]

the benefit of improved outlier performance. We observed A-CLAS to provide a more coherent multi-structure, multi-frame output, consistent with Wei et al.[8]

The advantage of A-CLAS over the frame-level segmentation is seen in the LV volume measurements and EF derived using the corresponding AP2 and AP4 views for each patient (Fig. 2). We report smaller mean absolute errors of 8.7ml (vs 9.9 ml[3]) in EDV, 6.3 ml (vs 6.6 ml) in ESV, and 4.6% vs (vs 5.3%) in EF. We also found smaller bias and narrower limits of agreement between the automated and reported clinical indices, with bias $\pm 1.96\sigma$ of 2.9mL $\pm$ 22.8 in EDV (vs 6.0mL $\pm$ 24.5[3]), 1.7mL $\pm$ 18.4 in ESV (vs 2.3mL $\pm$ 18.2), and 0.21% $\pm$ 12.0 in EF (vs 1.7% $\pm$ 14.3). These results are all within published inter-observer variability,[1] and represent the best results reported on this dataset.

The ground truth annotations of 50 test patients in CAMUS are intentionally left out for model evaluation through an online platform.[15] We generate the automated segmentation on the video clip of each test patient by aggregating the softmax outputs from all 10-fold models and then apply argmax across the channels. The results from the evaluation platform show A-CLAS achieved higher cross correlations on the EDV (0.983 vs 0.972), ESV (0.969 vs 0.963), and EF (0.883 vs 0.845) and consistent performance on segmentation compared with the best reported results on the platform.

We additionally assess the generalizability of CAMUS-trained A-CLAS on the large EchoNet-Dynamic clinical dataset from Stanford,[4] containing 1276 echocardiogram videos in AP4 view. In contrast to the single ED-ES video clips in the CAMUS dataset, the videos in EchoNet-dynamic dataset may contain multiple heartbeats. We extend Ouyang et al.[4] in using frame-level segmentation to determine all ED-ES clip locations in any video, and using the resulting multiple clips as test time augmentation. We generate the segmentations of all ED-ES video clips using the CAMUS-trained 10-fold models without tuning. The final estimation of a patient's EF is the average of derived EF from all clips. As shown in Figure 2, A-CLAS provides accurate EF estimation of 1.60% $\pm$ 15.95 over 1274 patients (two videos fail to return clips due to poor frame-level segmentation). These results are comparable to reported inter-observer variability.[1]

4. DISCUSSION

Our work here demonstrates the generalizability of temporally coherent multi-beat, multi-structure segmentation for large-scale echocardiography analysis. Our fair comparison between A-CLAS[8] joint video and motion tracking and frame-level segmentation methods[3] reveals A-CLAS's advantage in delivering the temporally consistent annotation for more accurate derivation of clinical indices. We report minimal bias and narrower limits of agreement on EDV, ESV, and EF with A-CLAS in both 10-fold cross-validation and in the held-out testing set of CAMUS. Assessing its generalizability to the EchoNet-Dynamic dataset validates A-CLAS's consistent performance on clinical data after training on the curated CAMUS set. Fully realizing the extension from 450 patients in CAMUS to the 10,030 patients (train, test, and validate sets) in EchoNet-Dynamic is a challenging task in transfer learning. We believe an effective domain adaptation[16] from a curated training dataset like CAMUS to often unlabeled clinical datasets will be the most crucial factor in this learning transformation.

REFERENCES

[1] Wood, P. W., Choy, J. B., et al., "Left ventricular ejection fraction and volumes: It depends on the imaging method," *Echocardiography* **31**(1), 87–100 (2014). https://doi.org/10.1111/echo.12331.

[2] Zhang, J., Gajjala, S., et al., "Fully automated echocardiogram interpretation in clinical practice," *Circulation* **136**(16), 1623–1635 (2018). https://doi.org/10.1161/CIRCULATIONAHA.118.034338.

[3] Stough, J. V., Raghunath, S., et al., "Left ventricular and atrial segmentation of 2d echocardiography with convolutional neural networks," in [*Medical Imaging 2020: Image Processing*], **11313**, 113130A, International Society for Optics and Photonics (2020). https://doi.org/10.1117/12.2547375.

[4] Ouyang, D., He, B., et al., "Video-based ai for beat-to-beat assessment of cardiac function," *Nature* **580**(7802), 252–256 (2020).

[5] Qin, C., Bai, W., et al., "Joint motion estimation and segmentation from undersampled cardiac mr image," in [*International Workshop on Machine Learning for Medical Image Reconstruction*], 55–63, Springer (2018).

[6] Qin, C., Bai, W., et al., "Joint learning of motion estimation and segmentation for cardiac mr image sequences," in [*Proc. Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*], (2018). https://doi.org/10.1007/978-3-030-00934-2_53.

[7] Li, M., Zhang, W., et al., "Recurrent aggregation learning for multi-view echocardiographic sequences segmentation," in [*Proc. Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*], (2019). https://arxiv.org/abs/1907.11292.

[8] Wei, H., Cao, H., et al., "Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape," in [*To appear, MICCAI*], (2020). https://www.researchgate.net/publication/342520911_Temporal-consistent_Segmentation_of_Echocardiography_with_Co-learning_from_Appearance_and_Shape.

[9] Çiçek, Ö., Abdulkadir, A., et al., "3d u-net: learning dense volumetric segmentation from sparse annotation," in [*Proc. MICCAI*], 424–432, Springer (2016).

[10] Leclerc, S., Smistad, E., et al., "Deep learning for segmentation using an open large-scale dataset in 2d echocardiography," *IEEE Trans Med Imaging* (2019). https://doi.org/10.1109/TMI.2019.2900516; https://www.creatis.insa-lyon.fr/Challenge/camus/index.html.

[11] Avants, B. B., Epstein, C. L., et al., "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Medical Image Analysis* **12**(1), 26–41 (2008).

[12] Balakrishnan, G., Zhao, A., et al., "Voxelmorph: a learning framework for deformable medical image registration," *IEEE Trans. Med. Imag* **38**(8), 1788–1800 (2019).

[13] Folland, E., Parisi, A., et al., "Assessment of left ventricular ejection fraction and volumes by real-time, two-dimensional echocardiography. a comparison of cineangiographic and radionuclide techniques," *Circulation* **60**, 760–766 (1979). https://doi.org/10.1161/01.cir.60.4.760.

[14] Pratt, J., "Remarks on zeros and ties in the wilcoxon signed rank procedures," *J. American Statistical Association* **54**, 655–667 (1959).

[15] Leclerc, S. et al., "Camus: Cardiac acquisitions for multi-structure ultrasound segmentation." https://www.creatis.insa-lyon.fr/Challenge/camus/index.html.

[16] Sankaranarayanan, S., Balaji, Y., et al., "Learning from synthetic data: Addressing domain shift for semantic segmentation," in [*Proc. CVPR*], (June 2018). https://openaccess.thecvf.com/content_ECCV_2018/papers/Yang_Zou_Unsupervised_Domain_Adaptation_ECCV_2018_paper.pdf.

5. SUPPLEMENTAL FIGURES

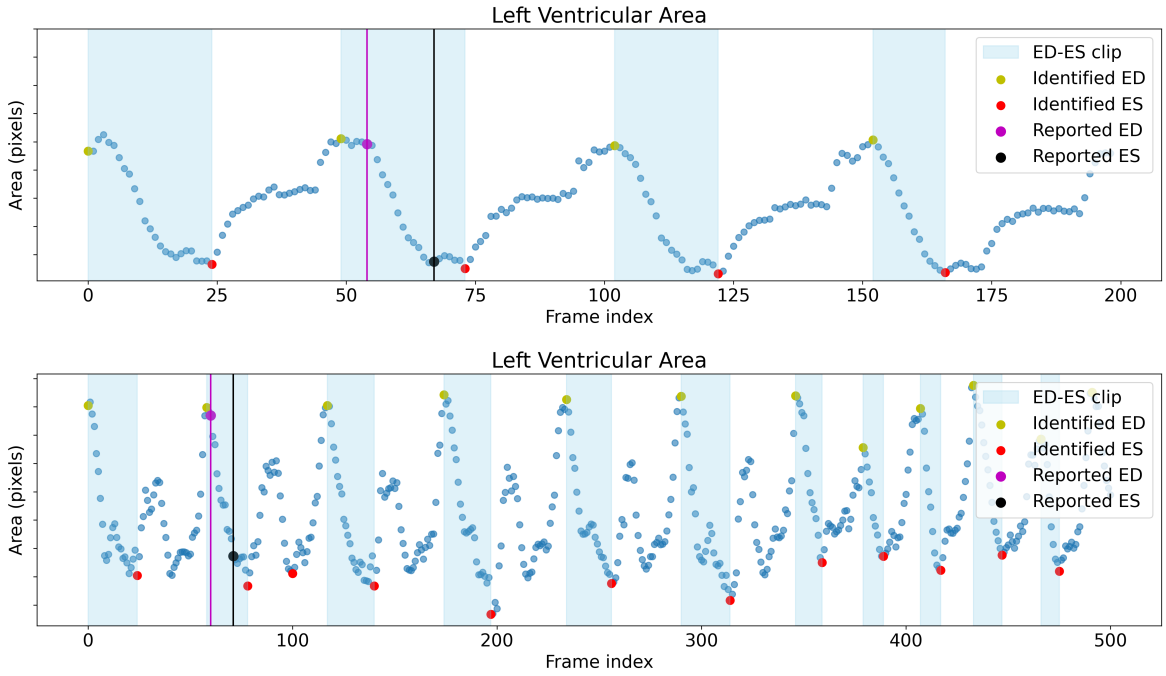

Figure 3. Clip finding of EchoNet-Dynamic videos using frame-level segmentation area. The vertical purple and black lines mark the clinically reported ED and ES frame of the video. The yellow and red dots mark the determined ED and ES frames, which are the peaks and troughs of the segmented left ventricular area (blue dots). The identified ED-ES video clips are shaded in blue.

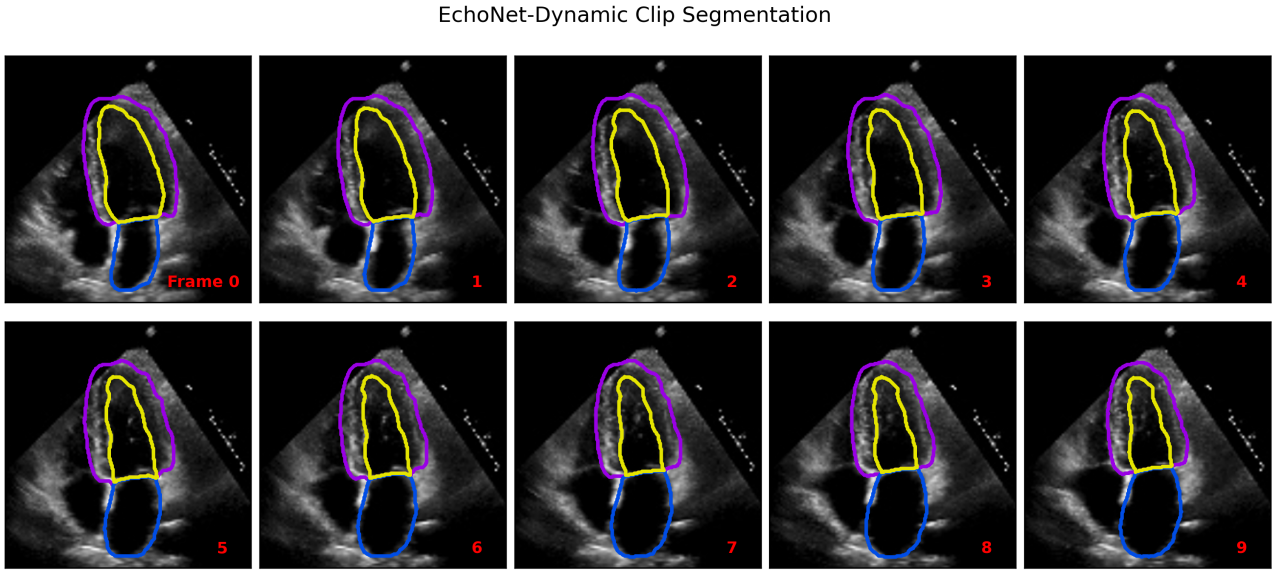

Figure 4. Segmentation output of an ED-ES (frame 0 to frame 9) video clip in the EchoNet-Dynamic dataset. The contours of the left ventricular epicardium, endocardium, and left atrium are delineated in purple, yellow, and blue respectively.

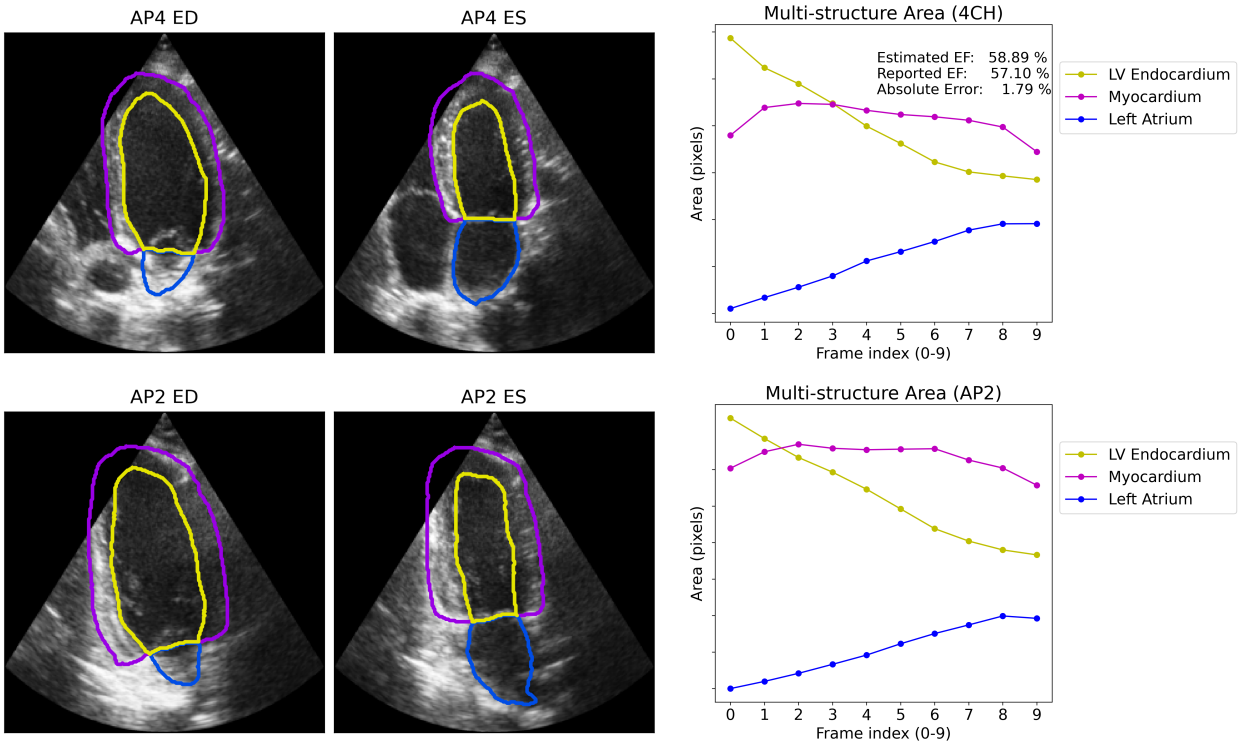

Figure 5. Example A-CLAS segmentation output in the CAMUS dataset. The right-most plots show the coherent change in area of the segmented structures, consistent with known cardiac motions.[8]