Fully Automated Multi-heartbeat Echocardiography Video Segmentation and Motion Tracking

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

Neural network-based video segmentation has proven effective in producing temporally-coherent segmentation and motion tracking of heart substructures in echocardiography. However, prior methods confine analysis to half-heartbeat systolic phase clips from end-diastole (ED) to end-systole (ES), requiring the specification of these frames in the video and limiting clinical applicability. Here we introduce CLAS-FV, a fully automated framework that extends upon this prior work, providing joint semantic segmentation and motion tracking in multi-beat echocardiograms. Our framework first employs a modified R2+1D ResNet stem, which is efficient in encoding spatiotemporal features, and further leverages sliding windows for both training and test time augmentation to accommodate the full cardiac cycle. First, through 10-fold cross-validation on the half-beat CAMUS dataset, we show that the R2+1D-based stem outperforms the prior 3D U-Net both in Dice overlap for all substructures, and in derived clinical indices of ED and ES ventricular volumes and ejection fraction (EF). Next, we use the large clinical EchoNet-Dynamic dataset to extend our framework to full multi-beat video segmentation. We obtain mean Dice overlap of 0.94/0.91 on left ventricle endocardium in ED/ES phases, and accurately infer EF (mean absolute error 5.3%) over 1269 test patients. The presented multi-heartbeat video segmentation framework promises fast and coherent segmentation and motion tracking for the rich phenotypic analysis of echocardiography.

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

1. INTRODUCTION

Echocardiography is a ubiquitous imaging modality for diagnosing and managing patients with cardiovascular disease.[1] Precise delineations of left ventricular endocardium (LV_{endo}) in echo support accurate derivation of ventricular volumes and ejection fraction (EF), which are important clinical indices. Cardiologists often inspect a multi-heartbeat echocardiogram and manually annotate the cardiac structures in one or more pairs of end-diastole (ED) and end-systole (ES) frames when inferring the patient's EF. Manual annotation in noisy ultrasonic images is labor-intensive and subject to high inter-rater variability.[2] Consequently, the segmentation of echocardiography using deep learning networks has been extensively studied to automate this process.

Promising echocardiography segmentation has been achieved by a variety of deep learning methods, including data-augmented frame-level segmentation,[3] frame-level segmentation guided by recurrent learning on spatiotemporal features,[4] and a combination of frame-level segmentation and a separate model for regressing on EF.[5] An emergent joint video segmentation and motion tracking network proposed by Wei et al.,[6] CLAS, recently proved superior in producing annotations that are consistent with cardiac motion.[7] CLAS improved the derived left ventricular end-diastolic volume (EDV), end-systolic volume (ESV), and EF estimation on the published CAMUS dataset[8] of echocardiograms in apical two (AP2) and four chamber (AP4) views.

However, when segmenting a typical clinically-acquired multi-beat echocardiogram, the framework would require a clinician's intervention or other out-of-band process (e.g, frame-level segmentation[7]) to identify each half-beat. Moreover, the CLAS framework confines its analysis to ED-to-ES half-heartbeat video clips provided by the CAMUS dataset. This latter requirement, in particular, relegates CLAS's applicability to systolic phase

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

analyses (i.e., the contractile phase of the heart cycle), precluding its ability to characterize diastolic function (i.e., the filling/relaxation phase of the heart cycle), which is highly relevant to many forms of heart disease including heart failure and valvular disease.[9]

We introduce CLAS-FV to completely automate full video, multi-beat echocardiogram segmentation, extending the CLAS framework with application to the large EchoNet-Dynamic dataset introduced by Ouyang et al.[5] CLAS-FV uses R2+1D spatiotemporal feature extraction, improving upon CLAS's 3D U-Net (see Sec. 2.1). We train the network using fixed-length clips that subsume the systolic phase, modifying the loss structure of CLAS to accommodate more of the cardiac cycle (Sec. 2.2). When segmenting an echocardiogram with multiple heartbeats, we divide the video into non-overlapping clips and then concatenate the results. We address potential discontinuities at clip boundaries through a limited sliding windows test time augmentation and label fusion (Sec. 2.3). Experimental results on both the CAMUS and Echonet-Dynamic are reported in Section 3.

2. ARCHITECTURE AND METHODS

In CLAS-FV, we use a shared feature extractor based on the R2+1D ResNet proposed by Tran et al.[10] for video action recognition. R2+1D ResNet has also been used for EF regression on multi-heartbeat echocardiogram in Ouyang et al.'s work.[5] The R2+1D convolutional block consists of a spatial 2D convolution followed by a temporal 1D convolution. This deconstruction accelerates the optimization and increases the network's nonlinearities,[10] which for us allow analyses of dynamic cardiac motion.

Our R2+1D-based shared feature extractor takes in an ordered sequence of echocardiogram frames and outputs a 64-channel feature map that has the same spatial and temporal shape as the input. To do this we capture the feature map outputs at all 5 R2+1D blocks along the ResNet encoder (output dimension $64, 64, 128, 256, 512$), upsampling and concatenating the feature maps before reducing the channel depth through successive 3D convolutions ($1024 \rightarrow 64 \rightarrow 64$, kernel size 1). The shared feature map is then sent to a segmentation head consisting of a 3D convolution layer and Softmax to acquire the putative segmentation of the input video over all frames (output dimension equal to the number of classes), and separately sent to a motion tracking head with a 3D convolution layer to derive bi-directional motion fields (output dimension 4); both task heads use kernel size 1.

2.1 R2+1D Versus 3D U-Net Feature Extraction

We validate the efficacy of the R2+1D-based shared feature extractor in producing both better segmentations and derived clinical indices, on the CAMUS dataset[8] containing systolic phase half-heartbeat echocardiograms. In a 10-fold cross-validation experiment, we compare R2+1D-based CLAS (∼32M parameters) to the original CLAS and its 3D U-Net[11] (∼19M parameters), and to a larger 3D UNet (∼34M parameters) where the intermediate feature map depths are modified to approximate the same number of trainable parameters. The networks are trained on sampled 10-frame ED-to-ES video clips with spatial size 128×128, intensity normalized to $[-1, 1]$.

Segmentation performance is measured by Dice overlap, $D(y_{auto}, y_{true}) = \frac{2(y_{auto} \cap y_{true})}{|y_{auto}|+|y_{true}|}$. We further derive the LV_{endo} volume through the Simpson's biplane method of disk,[12] which approximates the left ventricle as contiguous elliptical cylinders. The semi-major and semi-minor axes of elliptical cylinders are measured by the widths of segmented LV_{endo} in AP2 and AP4 views. The ejection fraction of a patient is the percentage change in LV_{endo} from ED to ES, $\frac{\text{EDV}-\text{ESV}}{\text{EDV}} \times 100\%$.

2.2 Full Video Segmentation and Motion Tracking

Beyond the updated feature stem, CLAS-FV learns full cardiac cycle segmentation through supervised learning on the EchoNet-Dynamic multi-beat dataset,[5] and through numerous changes to the loss structure associated with the motion tracking head. We train CLAS-FV with 32-frame video clips that subsume the manually denoted ED-to-ES sequence and its annotated boundary frames. Data augmentation is introduced by randomizing the clip start, such that training clips include more or fewer diastolic phase frames on either side of the denoted systolic phase sequence.

We modify the CLAS-FV loss structure to accommodate these variable starting points in the echo video sequence. Initially in Wei et al.'s CLAS framework, appearance level motion tracking is supervised by a combination of local cross correlation loss[13] and smoothness loss[14] called OTA. We replace OTA with a combined mean squared error loss and Huber loss,[15] which has proven effective in motion tracking in magnetic resonance.[16]

This change in appearance level motion tracking then affects the downstream losses in CLAS-FV. The output motion fields are used to spatially transform the two available labeled frames, from ED and ES fully forward in time to the end of the clip and fully backward to the beginning, to generate pseudolabels at all frames. Additional losses from Wei et al.[6] then compare these pseudolabels to the output of the segmentation head (SGS, binary cross-entropy) or to the ground truth labels themselves at the ES and ED frames respectively (OTS, multi-class dice). An additional binary cross-entropy loss trains the segmentation head separately (SGA).

The network is trained for 8 epochs using Adam optimizer with initial learning rate 1×10^{-4}. The learning rate is reduced to 1×10^{-5} after 3 epochs. The Huber loss is weighted by 0.005, and all other losses are equally weighted by 1. The model weights are saved with minimized loss on the validation set.

2.3 Test-time Augmentation and Clinical Indexing

When segmenting a multi-beat echocardiogram at test time, we divide the video into contiguous 32-frame clips, interpolating the video if necessary. On average, we obtain 5.6 32-frame clips from each test echocardiogram. The network segments all clips separately, and we concatenate the results to form a full video segmentation. Test time augmentation is used to improve coherence of resulting areas at the boundaries of these non-overlapping clips, repeating the segmentation process with four consecutive single-frame shifted versions of the video. We merge the segmentations at corresponding frames using the SIMPLE[18] label fusion technique.

To determine LV_{endo} volumes and EF, we use the area of the fused segmentation to identify all ED-ES systole phases (see Figure 1). With only AP4 views available in EchoNet-Dynamic, we compute the left-ventricular volume using Simpson's monoplane method that approximates the LV as contiguous circular cylindrical disks. We use the average derived EF from multiple systole phases as another test-time augmentation.

3. EXPERIMENTAL RESULTS

R2+1D Versus 3D U-Net: The CAMUS dataset contains 450 patients with echocardiograms in both AP2 and AP4 views (900 echocardiograms). Each video contains one half-heartbeat clip from ED to ES phase, and we resample to 10 frames with equal temporal interval. Manual annotations of left-ventricular endocardium (LV_{endo}), epicardium (LV_{epi}), and left-atrium (LA) are provided for the ED and ES frames. We perform a 10-fold cross-validation where each training patient appears in exactly one test fold. The sampling of patients is further stratified on EF range ($\leq 45\%, \geq 55\%, else$) and reported AP2 image quality (good, medium, poor) among folds, as suggested.[8]

As shown in Table 1, the R2+1D-based feature extractor achieves higher dice scores on all three substructures in both phases. A paired Wilcoxon signed-rank test[17] confirms the improvement of this feature extractor over both the original and larger 3D U-Net-based networks ($p \ll 0.001$). These improvements flow downstream to derived clinical indices as well, with R2+1D resulting in smaller mean absolute error (MAE) of 8.2 mL (vs 8.7 mL) in EDV estimation, 5.6 mL (vs 6.2 mL) in ESV, and 4.1% (vs 4.5%) in EF. The R2+1D-based CLAS takes ∼45 minutes for training of one test fold on Nvidia Titan RTX versus ∼70 minutes using the larger 3D U-Net, suggesting its faster optimization as another advantage.

Full Video Segmentation: EchoNet-Dynamic contains 10030 patient echocardiograms with one or more heartbeats in AP4 view. For each video, LV_{endo} is manually annotated by clinicians in one ED and one ES

Feature	Dice - ED			Dice - ES		
Extractors	$\text{LV}_{endo} \pm \sigma$	$\text{LV}_{epi} \pm \sigma$	$\text{LA} \pm \sigma$	$\text{LV}_{endo} \pm \sigma$	$\text{LV}_{epi} \pm \sigma$	$\text{LA} \pm \sigma$
3D U-Net[11]	0.938 ± 0.033	0.955 ± 0.022	0.874 ± 0.110	0.919 ± 0.045	0.949 ± 0.023	0.907 ± 0.079
Larger 3D U-Net	0.939 ± 0.035	0.955 ± 0.024	0.874 ± 0.101	0.918 ± 0.047	0.949 ± 0.025	0.907 ± 0.070
R2+1D-Based	0.944 ± 0.026	0.958 ± 0.018	0.884 ± 0.091	0.922 ± 0.045	0.952 ± 0.021	0.913 ± 0.066

Table 1. Mean dice overlaps (± standard deviation) on the multi-structural segmentation in ED and ES frames. The paired Wilcoxon signed-rank test[17] indicates that CLAS with R2+1D ResNet improves performance over CLAS with 3D UNet and Deeper 3D UNet on 450 training patients of CAMUS using 10-fold cross validation ($p \ll 0.001$).

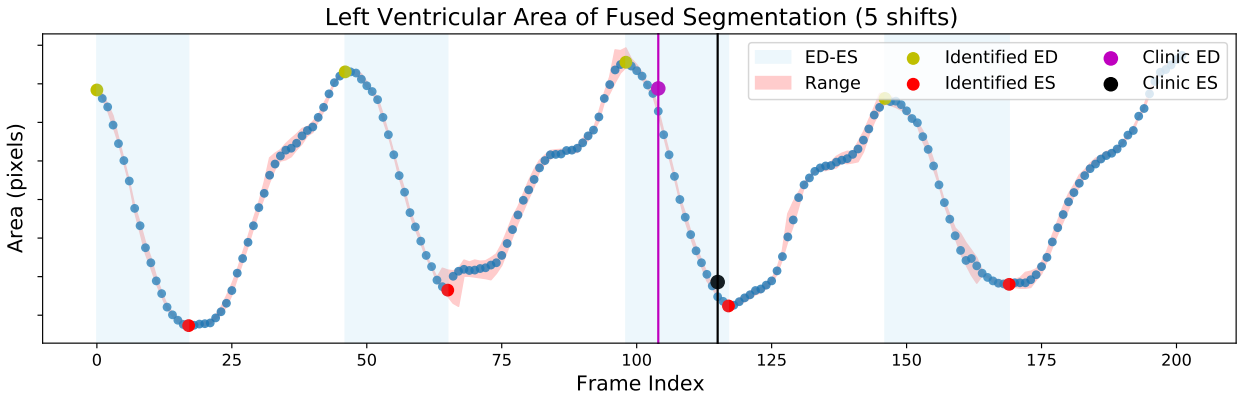

Figure 1. Identification of putative systolic phase clips of an EchoNet-Dynamic video using the fused LV_{endo} segmentation. The range of frame-level segmentation sizes on the original and 4 temporally shifted videos is shaded in red. The purple and black dots mark clinically reported ED and ES frames. The yellow and red dots mark the putative ED and ES frames from each cycle, which are the peaks and troughs of the fused segmented left ventricular area (blue dots). The identified systolic phase clips are shaded in blue.

frame. We use the Ouyang et al.'s split of train, validate, and test sets. However, we exclude from training and validation videos in which the ES frame precedes the ED frame or the denoted ED-ES clip is longer than 30 frames. As a result, CLAS-FV is trained on 7332 echocardograms and validated against 1258 that have the denoted systolic phase clip. We test the model with the best validation loss on 1276 test patients to assess its performance in LV_{endo} segmentation and EF estimation.

The test videos are segmented according to the shifting, splitting, and label fusion procedure outlined in Sec. 2.3. Once every frame is segmented, we can determine dice overlaps at the clinically denoted ED and ES frames. To determine EF given the full video segmentation, we use the peaks and troughs of the LV_{endo} size to determine systole phase clips, computing an EF for each and averaging the results over a video.

On 1276 test echocardiograms, CLAS-FV leads to LV_{endo} segmentation with mean dice overlaps of 0.935/0.907 in ED/ES frames. This compares favorably to prior CLAS-based assessments[7] while providing temporally coherent segmentation over all frames as opposed to only over systolic phase clips. For inference of EF, the automated workflow fails to identify systolic phase clips in 7 of 1276 test videos. Over the remaining 1269 patients, the MAE in derived EF versus clinically reported is 5.25%, compared to 5.83% for CLAS.[7] CLAS-FV also provides small bias and narrow limits of agreement, with bias $\pm 1.96\sigma$ of $-2.1\% \pm 12.9$, within inter-rater variability in non-contrast echocardiography.[2]

4. DISCUSSION

In this work we introduce CLAS-FV, a fully automated framework for the dense and temporally coherent segmentation of multi-beat echocardiograms. Building upon the prior CLAS model limited to systole phase analysis,[6] the CLAS-FV's convolutional network employs efficient spatiotemporal feature learning and losses appropriate for motion tracking to promote the co-learning of appearance and shape throughout the cardiac cycle. CLAS-FV further leverages sliding windows for both training and test time augmentation to accommodate the multiple cardiac cycles common to clinical echocardiography. We achieve state-of-art results on the extensive EchoNet-Dynamic dataset, for which previous analyses were limited to either non-coherent frame-level segmentation[5] or exclusively to the systole phase clips within the larger video.[7]

Moving forward, CLAS-FV offers the potential for detailed phenotypic analysis in large historical clinical datasets. Additinoally we look to study in particular the motion tracking outputs of CLAS-FV. Within the context of our test time augmentation and label fusion, these results are discarded. Recent work in motion tracking fusion may allow for additional small-scale heart motion analysis in clinical echo.

REFERENCES

[1] Virnig, B. A., Shippee, N. D., et al., "Trends in the use of echocardiography, 2007 to 2011." https://www.ncbi.nlm.nih.gov/books/NBK208663/.

[2] 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.

[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] 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.

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

[6] Wei, H., Cao, H., et al., "Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape," in [*MICCAI*], (2020). https://doi.org/10.1007/978-3-030-59713-9_60; https://www.researchgate.net/publication/342520911_Temporal-consistent_Segmentation_of_Echocardiography_with_Co-learning_from_Appearance_and_Shape.

[7] Chen, Y., Zhang, X., Haggerty, C. M., and Stough, J. V., "Assessing the generalizability of temporally coherent echocardiography video segmentation," in [*Medical Imaging 2021: Image Processing*], **11596**, 115961O, International Society for Optics and Photonics (2021).

[8] 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.

[9] Thomas, L., Marwick, T. H., Popescu, B. A., Donal, E., and Badano, L. P., "Left atrial structure and function, and left ventricular diastolic dysfunction: Jacc state-of-the-art review," *Journal of the American College of Cardiology* **73**(15), 1961–1977 (2019).

[10] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M., "A closer look at spatiotemporal convolutions for action recognition," in [*Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*], 6450–6459 (2018).

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

[12] 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.

[13] 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).

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

[15] Caballero, J., Ledig, C., Aitken, A., Acosta, A., Totz, J., Wang, Z., and Shi, W., "Real-time video super-resolution with spatio-temporal networks and motion compensation," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 4778–4787 (2017).

[16] 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).

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

[18] Langerak, T. R., van der Heide, U. A., Kotte, A. N., Viergever, M. A., Van Vulpen, M., and Pluim, J. P., "Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (simple)," *IEEE transactions on medical imaging* **29**(12), 2000–2008 (2010).

5. SUPPLEMENTAL FIGURES

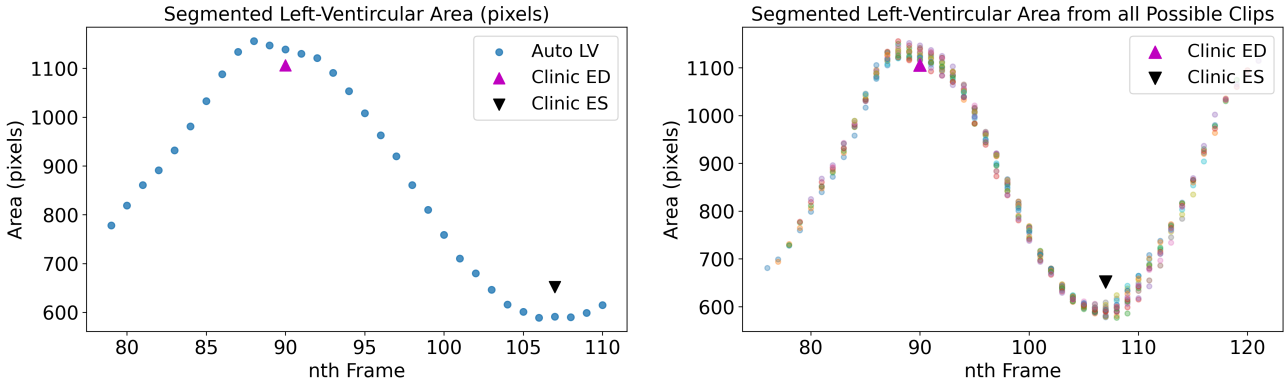

Figure 2. The left plot shows the size of segmented LV_{endo} (blue dots) in a 32-frame video clip of a training echocardiogram. The size of manual LV_{endo} labels at clinically identified ED and ES frames are marked in the purple and black triangles respectively. The 32-frame clip can start at an arbitrary time point as long as it covers the clinically identified ED-ES half-heartbeat. For the same echocardiogram, the right plot shows all possible 32-frame clips that can be chosen from the video. As data augmentation during training, one of the possible clips is randomly chosen during each epoch.

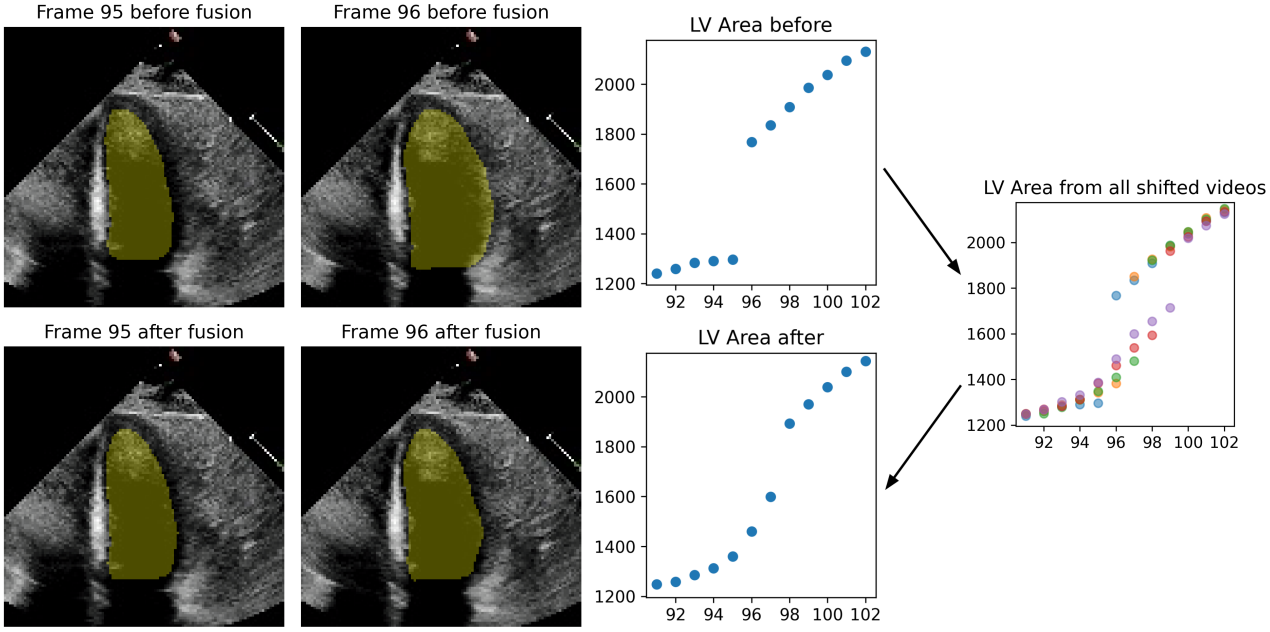

Figure 3. Segmented LV_{endo} at the boundary frames of two consecutive clips before and after label fusion[18] on the multi-segmentation of original and temporally shifted videos. The middle two scatter plots show the LV_{endo} area (pixels) around the boundary frames (frame 95 and 96) before and after fusion. The label fusion on segmentation of multiple temporally shifted clips connect the analysis that is previously independent between consecutive non-overlapping clips. As shown in plots, fused segmentation thus has a smoother change in LV area (pixels), consistent with known cardiac motion.

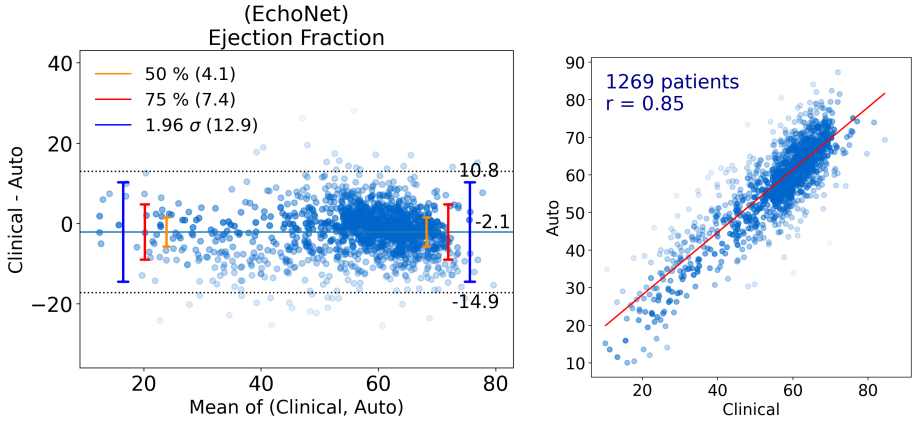

Figure 4. Bland-Altman plot of derived ejection fraction (EF) on EchoNet-Dynamic.[5] The blue solid line is the mean bias in EF estimation. The blue vertical brackets show bias $\pm 1.96\sigma$. The horizontal black dotted lines denote the inter-rater variability in non-contrast echocardiography.[2]