Temporally consistent segmentations from sparsely labeled echocardiograms using image registration for pseudo-label generation

Anonymous

Anonymous Organization
@*****.***

Abstract. The segmentation of the left ventricle in echocardiograms is crucial for the diagnosis of cardiovascular diseases. Deep learning-based methods have been shown to generate accurate segmentations reaching expert-level performance. However, these methods focus on accurate 2D segmentations of each frame while neglecting the temporal richness of the ultrasound sequence. This focus on 2D segmentation can be attributed to the scarcity of the manual annotations providing the training data, which are usually limited to the end-diastole and end-systole frames. This paper addresses these shortcomings and presents a novel method to train a temporally consistent segmentation model from sparsely labeled echocardiograms. We leverage an image registration-based approach to generate pseudo-labels for the frames that lack manual annotations, thereby creating a dataset that allows the training of 3D (2D+time) models. We evaluate our method through the segmentation of the left ventricle (LV) cavity, LV myocardium and left atrium using a state-of-the-art convolutional neural network (3D nnU-Net) on the public CAMUS dataset, and demonstrate the model's accuracy and temporal smoothness.

Keywords: Left ventricle segmentation · Echocardiography · Image registration · Pseudo-labels.

1 Introduction

The analysis of 2D transthoracic echocardiograms is ubiquitous in clinical cardiology, aiding in disease diagnosis and treatment selection [3]. Echocardiography features high spatial and temporal resolution, which makes it a key imaging tool for exposing structural and functional abnormalities of the heart. The analysis comprises the extraction of a number of quantitative markers of cardiac function, such as the ejection fraction (EF) and the chamber volumes [8]. Extraction of these quantitative markers requires accurate and precise delineation of the cardiac anatomy. However, manual expert annotation is a time-consuming task associated with high inter- and intra-rater variability [2]. Existing commercial solutions allow semi- or fully-automatic delineation of the cardiac structures, but they are typically limited to the segmentation of the end-diastolic (ED) and end-systolic (ES) frames [12].

The focus on ED and ES frames is also reflected in most machine learning approaches [10]. As these methods require large and diverse datasets for training, collecting annotations of full sequences has not been the prime focus: the largest and most commonly used public datasets for echocardiography segmentation, CAMUS [7] and EchoNet-Dynamic [9], only provide manual labels[1] for the ED and ES frames. Therefore, the current state-of-the-art (SoTA) segmentation methods rely exclusively on expert annotations for these two frames. Despite achieving performance within the margins of intra-observer variability [14][13], these methods do not address the smooth evolution of the cardiac structures over time, leading to temporally inconsistent predictions [10]. Since previous work showed that preserving the temporal consistency of the segmentations is beneficial for precise EF estimation [14], several recent studies have addressed this issue. For example, Li et al. proposed a method that exploits both temporal and multi-view information during training [1]. It employs a 3D CNN and a convolutional LSTM to respectively classify the sequence's view and output the 2D+time segmentation for the LV. Painchaud et al. imposed spatiotemporal smoothness during post-processing by using a constrained autoencoder to identify and correct spatiotemporal inconsistencies in segmented echocardiographic sequences [10]. Chen et al. finetuned their segmentation models using optical flow-adjusted annotations to improve motion coherence in the segmentation results, while also compensating for shape and size variations using a temporal affine network [4]. Wei et al. introduced an end-to-end approach that combines co-learning of appearance and shape features with the generation of LV pseudo-labels for the intermediate time points [14]. These LV pseudo-labels are obtained by warping the ground truth LV segmentation maps to other frames using optical flow. Although the method achieves temporally-consistent segmentation, its reliance on co-learning of appearance and shape features, combined with the need to generate the pseudo-labels, may result in a computationally complex algorithm. Moreover, the constrained end-to-end nature of the framework may limit its range of applications.

In contrast, we present a novel method that leverages an unsupervised image registration model to iteratively estimate the deformations between successive frames and generate pseudo-labels through the warping of the available segmentation maps. By combining sparsely annotated frames with pseudo-labeled intermediate frames, we create datasets that allow supervised training of arbitrary 3D (2D+time) segmentation networks. Our separation of the pseudo-label generation and segmentation reduces computational requirements compared with end-to-end approaches and unlocks potential for wider applicability. To this end, we train a 3D nnU-Net [6] to delineate the LV cavity, LV myocardium and left atrium (LA). We evaluate the proposed approach on the public CAMUS dataset [7], demonstrating that it leads to accurate segmentations that preserve spatiotemporal smoothnes and therefore yield accurate EF estimations.

[1] To aid readability, it may be worth specifying that "segmentations" and "labels" are used interchangeably throughout the paper.

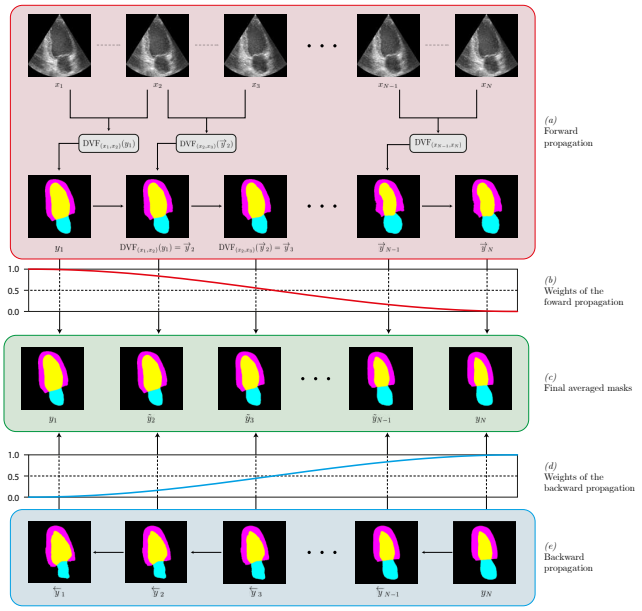

Fig. 1: The proposed DIR-based pseudo-labels generation method. The provided segmentations are propagated from ED to ES *(a)* and from ES to ED *(e)*. The masks from the two directions are subsequently aggregated as described in Section 2.1 and weighted according to a sinusoidal function *(b and d)*.

2 Method

To obtain accurate and temporally consistent 3D (2D+time) segmentations from a sparsely labeled dataset, the method first generates the pseudo-labels for those frames that lack reference segmentations. This is done through the iterative application of image registration. Thereafter, the method uses these pseudo-labels to augment sparse reference annotations and train a segmentation model.

2.1 Pseudo-labels generation

Echocardiography acquisition consists of a sequence of image frames x_t, $\forall t \in \{1, 2, .., N\}$ showing the evolution of the heart over the cardiac cycle. Given the reference segmentation for the ED and ES frames, unsupervised deformable image registration (DIR) is exploited to segment the frames lacking segmentation masks. The registration's dense displacement vector field (DVF) is employed to warp the segmentation of frame x_t (y_t) to frame x_{t+1}, resulting in a pseudo-segmentation $\overrightarrow{y}_{t+1}$ of frame x_{t+1}. Specifically, the available ED segmentation is iteratively forward-propagated through the sequence to produce $\overrightarrow{y}_t$, $\forall t \in \{1, 2, .., N\}$. Akin, backward-propagating the ES segmentation mask returns a set of $\overleftarrow{y}_t$, $\forall t \in \{1, 2, .., N\}$.

Given that sequential registrations may lead to error accumulation, the two sets of pseudo-labels $\overleftarrow{\mathbf{y}}$ and $\overrightarrow{\mathbf{y}}$ are averaged by weighting each frame with the distance from the ED and ES reference segmentations, respectively. For this, class-wise binary masks are extracted and the signed distance to their edges is calculated. Weighted-averaging the signed distances returns an image where pixels outside the object have negative values, inner pixels have positive values and the object boundaries are located at the zero crossings. Therefore, thresholding this image at zero yields the final mask. The final *bidirectional* method is schematized in Figure 1 and mathematically described in Equation 1:

$$\tilde{y}_{t,C} = d(\overrightarrow{y}_{t,C}) \cdot \cos^2 \frac{\pi}{2N} t + d(\overleftarrow{y}_{t,C}) \cdot \sin^2 \frac{\pi}{2N} t > 0 \tag{1}$$

where $\overleftrightarrow{y}_{t,C}$ is the binary mask corresponding to class C at time point t, $d(\cdot)$ is the distance transform operation and N is the sequence length. The sinusoidal weights are designed to decrease from 1 to 0 in the direction of the propagation, thereby exerting more influence on the forward direction at the beginning of the sequence and on the backward direction at the end. This further mitigates error accumulation and improves the accuracy of the object representation.

In this work, an unsupervised deep learning registration framework is utilized to perform image alignment through CNNs [5]. The method exploits image similarity between fixed and moving image pairs, B-splines as the transformation model, and supports coarse-to-fine alignment. Additionally, the loss function combines the negative normalized cross correlation $\mathcal{L}_{NCC}$ with the bending energy penalty P: $\mathcal{L} = \mathcal{L}_{NCC} + \alpha P$ [11]. The regularization term P minimizes the second order derivative of local transformations, thereby enforcing global smoothness and preventing anatomically implausible image folding.

2.2 Segmentation

The reference segmentations of the echocardiograms are augmented with the pseudo-labels to provide densely labeled reference sequences. This enables the training of 3D (2D+time) segmentation models, which are designed to be trained on densely annotated data. By encoding the time dimension as the third dimension in convolutional space, a 3D model can learn spatiotemporal features that encourage temporally smooth predictions. To this end, a 3D nnU-Net is trained on the augmented dataset (*3D Dense* nnU-Net)[6].

3 Experiments

Two main experiments were conducted[2]. First, the pseudo-labels were generated and evaluated against reference segmentations. Second, the pseudo-labels were utilized to complement the original dataset and train the segmentation network.

All the models were implemented in PyTorch 1.12.1 and trained using 2 Intel Xeon Gold 6128 CPUs (6 cores, 3.40GHz) and a GeForce RTX 2080 Ti.

[2] The code is publicly available at https://anonymous.4open.science/r/miccai-temporally-consistent-echosegmentation.

3.1 Data and preprocessing

This study uses two public datasets: CAMUS [7] and TED [10]. CAMUS contains 2D echocardiograms with 2-chambers (2CH) and 4-chambers (4CH) views of half-cycle sequences (from ED to ES) of 500 patients (450 training, 50 test). Manual annotations of the LV cavity, LV myocardium and LA are provided for the ED and ES frames only. TED is a subset of CAMUS that comprises 98 full cycle 4CH sequences with manual segmentations of the LV cavity and the LV myocardium for the *whole* cardiac cycle. 94 sequences are part of the CAMUS training set and 4 of the test set.

Throughout this work, all images are resized to 512×512 px, and the pixel spacing is scaled proportionally to preserve the anisotropic nature of the data.

3.2 Evaluation Metrics

Both the generated pseudo-label segmentations and the predicted segmentations are intrinsically evaluated by overlap and boundary metrics: the DICE coefficient (DC), the mean absolute surface distance (MAD) and the 2D Hausdorff Distance (HD). Since the metrics are calculated per frame, they are averaged over an entire video. Additionally, the segmentation models are evaluated extrinsically on the EF and LV volumes (at end-diastole and end-systole, EDV and ESV) estimation. To aggregate dataset-level statistics for these estimations, the correlation coefficient, bias and mean absolute error (MAE) are calculated between the reference and automatically obtained values. Finally, the temporal consistency of the automatic segmentation is assessed by tracking the area of a given class over time. The smoothness of a sequence is computed as the integral of the second derivative of the resulting curve, which will be denoted as *area curve*. To account for changes in the slope of the area curve and prevent the loss of information due to opposite bending, the second derivative is squared prior to integration. The final smoothness metric is defined in Equation 2, with N being the sequence length and $a_C(t)$ the area of class C at time point t.

$$\text{Smoothness} = \int_1^N \left(a_C''(t)\right)^2 dt, \tag{2}$$

3.3 Pseudo-label Generation

The DIR model was trained on the CAMUS training set after leaving out the overlapping 94 TED echocardiograms, resulting in a set of 806 echo sequences. Successively, the frame-wise alignment quality was evaluated against these 94 left-out TED sequences. The DIR network was trained on every possible intra-patient combination of two frames from the registration training set. The training was performed in 10,000 iterations and used a batch size of 32, the AMSGrad variant of the ADAM optimizer and a learning rate of 10^{-3}. Hyperparameters such as the number of kernels, the kernel size and the B-spline grid spacing were determined in preliminary experiments by testing values between 2 and 128. The

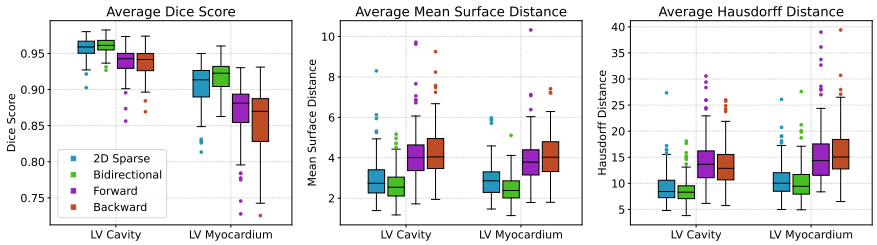

Fig. 2: Comparison of the pseudo-labels quality in terms of geometric metrics.

best results were achieved using 32 kernels of size 32×32, a grid spacing of 32 and a regularization hyperparameter value of 1.0 to prevent folding. No performance gain was observed when performing coarse-to-fine registration, therefore simple one-stage alignment was utilized.

Figure 2 shows the performance of the pseudo-label generation using the forward, backward and bidirectional approach. For comparison, pseudo-labels were also generated using a SoTA 2D nnU-Net trained on the original sparsely labeled CAMUS dataset (*2D Sparse nnU-Net*), to allow a comparison between our DIR-based pseudo-labels and the predictions of a segmentation model.

3.4 Segmentation

The *3D Dense* nnU-Net was trained and tested on the sparsely labeled CAMUS datasets augmented with pseudo-labels, allowing direct comparison with related works. In addition, the *3D Dense* model was evaluated against two baselines: a 2D nnU-Net trained on the sparsely labeled CAMUS dataset (*2D sparse* nnU-Net) and a 2D nnU-Net trained on the *augmented* CAMUS dataset (*2D Dense* nnU-Net). Each nnU-Net was trained for 1,000 epochs, using 5-fold cross-validation with an interleaved test setup. After training, the framework automatically selected the best U-Net configuration. Finally, the SoTA method by Wei et al., CLAS, is included in the comparison [14]. The models were compared in terms of (i) accuracy of the LV cavity, LV myocardium and LA segmentation at ED and ES; (ii) estimation of EF, EDV, and ESV; (iii) temporal smoothness.

The average segmentation performance on the ED and ES frames of the test set is listed in Table 1; the results of the EDV, ESV and EF estimation are displayed in Table 2; the observed temporal consistency of frame-by-frame predictions is shown in Figure 3; finally, the area curve of a test patient is depicted in Fig. 4 along with the corresponding ED and ES predictions.

4 Discussion and conclusion

This paper presented a novel method for obtaining temporally consistent segmentations of echocardiography using sparsely labeled data. The method exploits pseudo-labels generated by the use of DIR to complement the original set of sparsely annotated frames and allow the training of a 3D nnU-Net.

ED	LV Cavity			LV Myocardium			LA		
	DC	HD	MAD	DC	HD	MAD	DC	HD	MAD
Intra-observer	0.945	4.6	1.4	0.957	5.0	1.7	–	–	–
CLAS [14]	0.947	4.6	1.4	0.961	4.8	1.5	0.902	5.2	**1.9**
2D Sparse	**0.955**	**4.1**	**1.2**	**0.965**	4.4	**1.4**	**0.906**	**4.9**	**1.9**
2D Dense	0.950	4.2	1.3	0.963	**4.3**	**1.4**	0.902	5.0	2.0
3D (2D+time)	0.952	4.2	1.3	0.961	4.6	1.5	0.899	5.2	2.0

ES	LV Cavity			LV Myocardium			LA		
	DC	HD	MAD	DC	HD	MAD	DC	HD	MAD
Intra-observer	0.930	4.5	1.3	0.951	5.0	1.7	–	–	–
CLAS [14]	0.929	4.6	1.4	0.955	4.9	1.6	0.927	4.8	1.8
2D Sparse	0.938	**4.0**	**1.2**	**0.959**	**4.3**	**1.5**	**0.937**	**4.3**	**1.5**
2D Dense	0.934	4.2	1.3	0.957	4.5	**1.5**	0.933	4.5	1.7
3D (2D+time)	**0.939**	**4.0**	**1.2**	0.958	4.8	**1.5**	0.932	4.7	1.6

Table 1: Average segmentation results at ED and ES. The intra-observer variability results (in red) are taken from the official CAMUS website and are not provided for the left atrium. The best value per column is indicated in bold.

Methods	EDV			ESV			EF		
	Corr	Bias	MAE	Corr	Bias	MAE	Corr	Bias	MAE
Intra-observer	0.978	-2.8	6.5	0.981	-0.1	4.5	0.896	-2.3	4.7
CLAS [14]	0.958	-0.7	7.7	0.979	**-0.0**	4.4	**0.926**	**-0.1**	**4.0**
2D Sparse	0.972	**0.0**	6.0	0.98	-0.6	4.8	0.827	1.3	5.0
2D Dense	0.972	0.4	5.7	**0.986**	-0.3	4.2	0.841	1.3	4.6
3D (2D+time)	**0.978**	-1.4	**4.8**	**0.986**	-0.1	**4.0**	0.859	**-0.1**	4.6

Table 2: LV volume and EF estimation on the test set. The intra-observer variability is indicated in red, and the best column-wise value is displayed in bold.

The analysis of the pseudo-label quality revealed the benefits of bidirectional over unidirectional label propagation. The results on the ED and ES segmentation tasks demonstrate that exploiting the pseudo-labels retains or improves the performance of the model trained on the sparsely labeled dataset, thereby endorsing their quality for downstream applications. In fact, the overlap and boundary metrics show that all three evaluated models perform *at least* as well as the SoTA CLAS method, and achieve a level of accuracy on par with intra-observer variability. Evaluation of the temporal smoothness of the segmentation showed that the *2D Dense* model outperforms the *2D Sparse* model and that the 3D model, in turn, outperforms both.

The same pattern is observed in the LV volumes and EF estimation, supporting the claim that enforcing temporal consistency in the segmentations is beneficial for EF estimation. Note that the *3D Dense* method computes strong EDV and ESV estimates that are well within the intra-observer variability and superior to CLAS. The model's estimation of the EF, however, is slightly less remarkable. Still, we argue that the very low bias and the MAE akin to intra-rater precision advocate sufficiently good estimations of the measure. Additionally, it is noteworthy that accurate EF estimation is only one of the practical benefits of temporal smoothness. For instance, temporally consistent LV segmentations can help to track the motion of the ventricular walls, which is essential for calculating

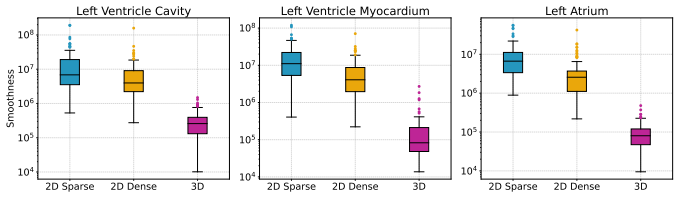

Fig. 3: Temporal smoothness of the test set predictions in terms of the metric defined in Equation 2 (lower values indicate higher smoothness). Note the logarithmic scale on the y-axis.

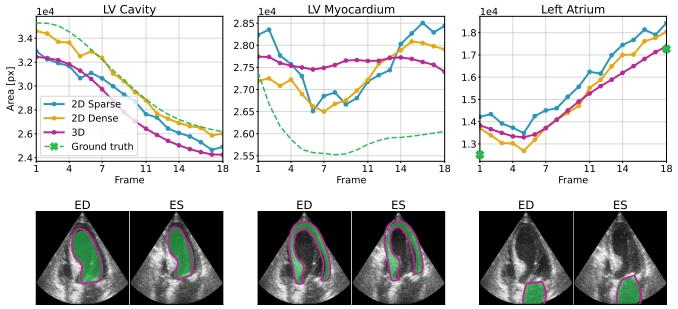

Fig. 4: Evaluation of the temporal consistency on `patient0002` from the test set. *Top row:* area curves. *Bottom row:* corresponding predictions at ED and ES. Following the legend of the curves, the green area refers to the ground truth and the magenta outline is the prediction of the *3D Dense* model.

strain and other measures of cardiac function. Similarly, smooth segmentations are crucial for recognizing functional abnormalities [9].

In Figure 4 the 3D approach appears to be offset from the ground truth and the 2D models, especially at ED and ES. Inspection of other patients revealed that the model is not biased towards over- or under-segmentation of the structures. Rather, Fig. 4 suggests the presence of aleatorically uncertain boundaries in the data. There are clear disagreements between the manual and automatic segmentations of the LV myocardium due to the occluded endocardium. An analysis of other test patients unveiled similar discrepancies when the LV myocardium and/or the LA extend beyond the field of view. In these cases, the ambiguous position of the structures is presumably reflected in the creation of the manual annotations. Accordingly, the predictions of our models contain uncertainty, resulting in the observed discrepancy. Future work could model this aleatoric randomness in order to convey the reliability of a given estimation.

To conclude, the proposed approach successfully enabled the creation of an accurate and temporally consistent segmentation model. The flexible and modular nature of our method enables the use of different architectures, thereby promoting future developments.

References

1. Mv-ran: Multiview recurrent aggregation network for echocardiographic sequences segmentation and full cardiac cycle analysis. Computers in Biology and Medicine **120**, 103728 (2020). https://doi.org/https://doi.org/10.1016/j.compbiomed.2020.103728
2. Armstrong, A.C., Ricketts, E.P., Cox, C., Adler, P., Arynchyn, A., Liu, K., Stengel, E., Sidney, S., Lewis, C.E., Schreiner, P.J., Shikany, J.M., Keck, K., Merlo, J., Gidding, S.S., Lima, J.A.C.: Quality control and reproducibility in m-mode, two-dimensional, and speckle tracking echocardiography acquisition and analysis: The CARDIA study, year 25 examination experience. Echocardiography **32**(8), 1233–1240 (Nov 2014). https://doi.org/10.1111/echo.12832
3. Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., Rueckert, D.: Deep learning for cardiac image segmentation: A review. Frontiers in Cardiovascular Medicine **7** (Mar 2020). https://doi.org/10.3389/fcvm.2020.00025
4. Chen, S., Ma, K., Zheng, Y.: Tan: Temporal affine network for real-time left ventricle anatomical structure analysis based on 2d ultrasound videos. ArXiv (2019). https://doi.org/10.48550/ARXIV.1904.00631
5. de Vos, B.D., Berendsen, F.F., Viergever, M.A., Sokooti, H., Staring, M., Išgum, I.: A deep learning framework for unsupervised affine and deformable image registration. Medical Image Analysis **52**, 128–143 (2019). https://doi.org/10.1016/j.media.2018.11.010
6. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (Dec 2020). https://doi.org/10.1038/s41592-020-01008-z
7. Leclerc, S., Smistad, E., Pedrosa, J., Ostvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., Lartizien, C., Dhooge, J., Lovstakken, L., Bernard, O.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Transactions on Medical Imaging **38**(9), 2198–2210 (Sep 2019). https://doi.org/10.1109/tmi.2019.2900516
8. Moal, O., Roger, E., Lamouroux, A., Younes, C., Bonnet, G., Moal, B., Lafitte, S.: Explicit and automatic ejection fraction assessment on 2d cardiac ultrasound with a deep learning-based approach. Computers in Biology and Medicine **146**, 105637 (2022). https://doi.org/https://doi.org/10.1016/j.compbiomed.2022.105637
9. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., Zou, J.Y.: Video-based AI for beat-to-beat assessment of cardiac function. Nature **580**(7802), 252–256 (Mar 2020). https://doi.org/10.1038/s41586-020-2145-8
10. Painchaud, N., Duchateau, N., Bernard, O., Jodoin, P.M.: Echocardiography Segmentation with Enforced Temporal Consistency. IEEE Transactions on Medical Imaging **41**(10), 2867–2878 (Oct 2022). https://doi.org/10.1109/TMI.2022.3173669
11. Rueckert, D.: Nonrigid registration using free-form deformations: Application to breast mr images. IEEE Transactions on Medical Imaging **18**(8), 712 – 721 (1999). https://doi.org/10.1109/42.796284
12. Schuuring, M.J., Išgum, I., Cosyns, B., Chamuleau, S.A.J., Bouma, B.J.: Routine echocardiography and artificial intelligence solutions. Front. Cardiovasc. Med. **8**, 648877 (Feb 2021)

13. Sfakianakis, C., Simantiris, G., Tziritas, G.: Gudu: Geometrically-constrained ultrasound data augmentation in u-net for echocardiography semantic segmentation. Biomedical Signal Processing and Control **82**, 104557 (2023). https://doi.org/https://doi.org/10.1016/j.bspc.2022.104557
14. Wei, H., Cao, H., Cao, Y., Zhou, Y., Xue, W., Ni, D., Li, S.: Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pp. 623–632. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-59713-9$_6$0