Project Final Report- 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

This project is dedicated to delivering swift and comprehensive solutions for echocardiography analysis. Specifically, our goal is to construct a model for left ventricle segmentation and ejection fraction estimation using video analysis techniques learned in the class. While previous studies have extensively covered these topics, most have relied on image analysis rather than video. The limitation of image-based methods lies in their focus on key frames, typically the last frames during systolic and diastolic phases. This approach misses the opportunity for temporal tracking of the left ventricle and overlooks valuable information within the frames in-between. In this project, our aim is to develop a video-based model utilizing complete, uncurated videos to provide simultaneous temporal segmentation for the left ventricle and estimation for ejection fraction.

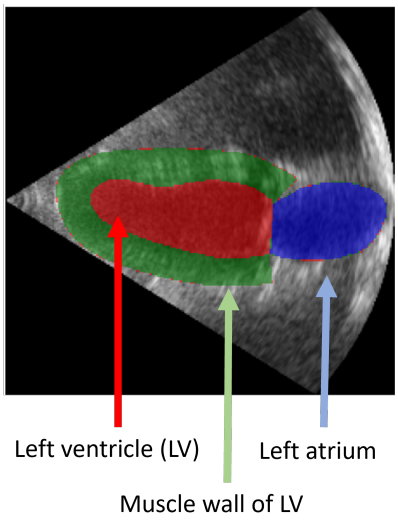

Figure 1. Segmentation targets of the heart

1. Problem Definition

1.1. Echocardiography

Echocardiography, often referred to as "heart ultrasound," is a valuable medical imaging technique for assessing heart structure and function. Documented in video format, it aids in early heart problem detection, monitors cardiac conditions, and assesses the effectiveness of medical interventions. However, echocardiography encounters two primary limitations: the video quality and its interpretations are highly operator-dependent, often demanding a significant amount of time from 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) temporal segmentation and ejection fraction estimation (a regression task).

1.2. Left Ventricle Segmentation

The left ventricle (LV) is the major chamber within the heart, as indicated by the red arrow in Fig 1. It is linked to the left atrium (LA) responsible for propelling blood into the LV prior to its contraction. In clinical settings, the segmentation of the LV, particularly temporal segmentation, proves tools for physicians in assessing the heart's contraction pattern. This segmentation not only helps thoroughly assess the structure and function of the heart but also provides valuable insights for diagnosing major heart diseases. Since LV is the main engine of the heart, it is the key factor for determining the ejection fraction.

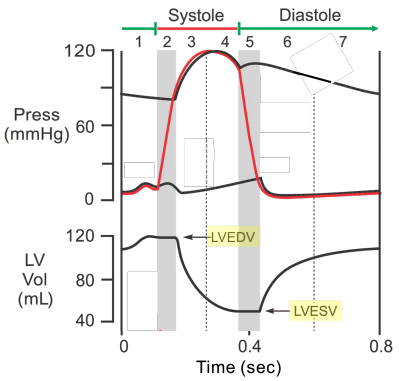

Figure 2. A single heart cycle composing of systolic and diastolic phases.

1.3. Ejection Fraction Estimation

Ejection fraction estimates (EF estimation) 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:

$$EF = \frac{EDV - ESV}{EDV} \times 100\% \quad (1)$$

where

- EDV stands for End-Diastolic Volume, indicating the commencement of heart contraction. At this juncture, the LV is in a fully relaxed state and possesses its maximum volume, filled with blood.
- ESV stands for End-Systolic Volume, marking the conclusion of heart contraction. During this moment, the LV has expelled the maximum amount of blood it can pump.

Figure 2 illustrates the volume-pressure relationship of the LV during the systolic phase.

1.4. Overall Goals

Our main goal is to provide temporal LV segmentation with EF estimation in a single step. As mentioned earlier, the LV acts as the engine driving heart contraction. Therefore, we believe that observing the 2D projection of the LV indicates changes in its volume, and thus, can be used to predict ejection fraction.

Unlike previous state-of-the-art models that often treat ED and ES (the first and last frames of the contraction) as two independent frames and train image-based models on them, we aim to train a video-based model that correlate ED, ES, and all the frames in-between temporally. Also, we'd like to provide segmentation masks for all the frames in-between, instead of simply focusing on ED and ES like the before studies. Therefore, the input to our model will be short clips of heart ultrasound videos. These video clips can contain either a single cycle of heart beating or multiple cycles.

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, there are no previous model considering the power of any state-of-the-art 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].

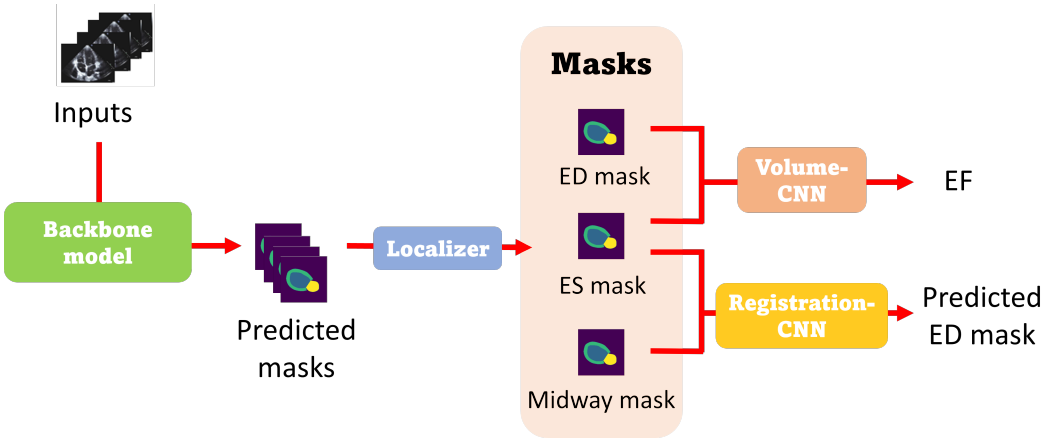

Figure 3. The overview of our model framework involves completing three sub-tasks. The backbone model predicts temporal masks for the left ventricle. A rule-based localizer then helps pick out masks for the first frame (ED), the midway frame, and the last frame (ES) from the predictions. A downstream CNN takes ED and ES masks as inputs to estimate ejection fraction (EF). Additionally, the registration CNN uses ED and midway masks as input to predict the ES mask. We aim to use either DeepLabv3 or Mask2Former as the backbone model.

3. Technical Approach

3.1. Architectures Overview

Our model is composed of three sub-modules: a backbone model that predicts temporal segmentation masks, a CNN-based model that takes segmentation masks to predict ejection fraction, and a registration model that provides temporal regularization for mask prediction.

Note that TimeSformer[8] and R-CNN[9] were once considered as backbone in our first proposal. However, due to lacking the decoder part for reconstructing segmentation masks, we did not successfully implement them for our objectives.

3.2. Backbone Model

The backbone model is responsible for offering temporal segmentation for the input videos. In this project, we investigate DeepLabv3 [10] and Mask2Former [11] as our backbone models. Both DeepLabv3 and Mask2Former were initially designed for image-based segmentation. To enable these models to capture the temporal relationship between frames, we introduce a temporal loss to constrain the mask predictions.

DeepLabv3: 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.

Mask2Former: We also opted for Mask2Former, a cutting-edge model for universal image segmentation published by META in 2022, to represent transformer models and assess performance against CNN-based models. Mask2Former utilizes a masked attention paradigm, eliminating the need for explicit object queries or proposals seen in R-CNN, thereby significantly simplifying the data flow and improving efficiency. Moreover, Mask2Former harnesses the strengths of the Swin Transformer [12] to extract rich features from the input image. This combination of masked attention and Swin Transformer empowers Mask2Former to achieve state-of-the-art results on various segmentation benchmarks.

Below, we emphasize three fundamental innovations inside this model:

- It exploits a novel Transformer decoder, different from conventional models, capable of using the mask attention mechanism for finding localized features.
- Through a Multi-scale strategy, it is also able to handle small objects effectively, harnessing high-resolution features and with efficient computational demands.
- Mask2Former strategically switches the order of self- and cross-attention, optimizing computational efficiency.

After training, Mask2Former takes an input image, creates per-pixel embeddings in a multi-resolution pyramid, and provides binary segmentation masks for all the possible objects seen in the image.

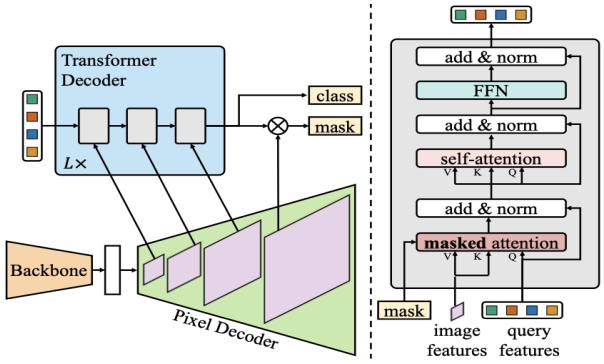

Figure 4. Mask2former architecture.

3.3. Temporal Geometric Constraint For Heart Masks

Since the heart beats steadily, ensuring the temporal smooth 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 the predicted masks. A Temporal Smoothness Loss (TS loss) penalizes deviations in intensity between adjacent masks.

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

Differing from the conventional smoothness loss computed on entire images, our TS loss is calculated on binary masks individually for each class. Specifically, we compute the TS loss for the two heart chambers, LV and LA, seen in the image.

3.4. Rule-based Frame Localizer

For the computation of registration and EF, knowing the locations of ED and ES frames is essential. To address this, we employ a small rule-based selection module. We make the assumption that, for a significant portion of patients in the training set, the ED frame corresponds to the frame containing the largest segmentation mask, while the ES frame corresponds to the frame containing the smallest segmentation mask. We acknowledge that this assumption may be challenged by variations in the angle of the heart ultrasound and the 2D projection of the heart volume. However, we believe that the benefits offered by the proposed rule-based localizer outweigh potential noise from these factors.

3.5. Mask-based Volume Estimation

Most prior models directly estimated ED and ES volumes from their respective single-frame images. They then follow Equation 1 to compute the final EF. This approach requires manual annotation for the locations of ED and ES and does not consider the possible temporal relationship between them. Importantly, using the entire image as input might be susceptible to background noise.

Contrast to the previous approaches, our model estimates EF based on the predicted segmentation masks instead of the entire image. Specifically, the steps are listed as:

- Our model takes the temporal segmentation masks of LV as inputs and estimates the LV volume for each frame.
- Similar to temporal segmentation, we compute temporal smoothness for the volume prediction for the adjacent frames.
- The rule-based localizer assists in selecting the frame ID for both ED and ES. We then utilize their corresponding predicted volumes to calculate the final EF.

3.6. Mask Registration

Another innovation of our approach is imposing optical flow registration on predicted masks. We assume that the speed of the optical flow for each pixel in the LV segmentation is relatively stable and predictable based on its past locations. Therefore, we treat this future mask prediction as a registration problem. Specifically, our registration module takes the first predicted frame (ED) and the middle predicted frame (midway) to predict a deformation field. Below, we propose two different approaches for applying the predicted deformation field. The major difference between this two approaches is the receiver of the deformation field. The first approach applies the deformation field to the first frame to transform it into the predicted last-frame mask. The second approach applies the predicted deformation field to the midway mask to obtain the predicted last-frame mask. In the end, these predicted last-frame mask is used to compute the L2 loss with the predicted last-frame mask from the backbone model to regularize its output. Fig 5 illustrates the two proposed approaches. In our study, we used UNet to realize the deformation field prediction.

4. Experiments

4.1. Evaluation

LV segmentation:

we evaluate our segmentation results on the temporal Dice [13]. The Dice coefficient, a widely used metric for evaluating semantic segmentation tasks, measures the spatial overlap between predicted and ground-truth segmentation masks. In video segmentation, the temporal Dice coefficient extends the traditional Dice coefficient by incorporating temporal information into the calculation. Mathematically, this can be expressed as

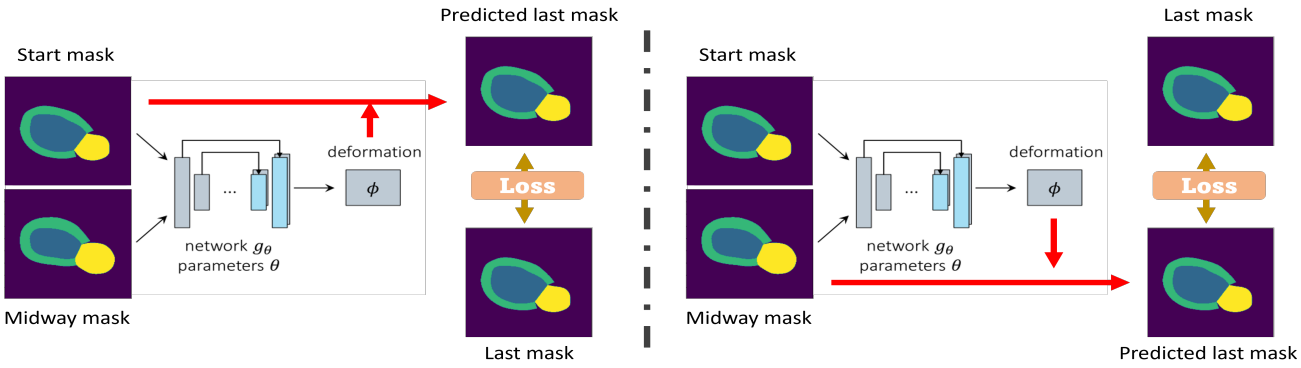

Figure 5.

$$\text{t-Dice} = \frac{\sum_{t=1}^{T} 2V_{overlap}^{t}}{V_{pred}^{t} + V_{gt}^{t}} \tag{3}$$

where $V_{overlap}^{t}$ is the overlap between the predicted and ground truth masks at frame t, and V_{pred}^{t} and V_{gt}^{t} are the predicted and ground truth masks at frame t, respectively.

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

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 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 established in 2021 to facilitate a comprehensive study of cardiac motion and chamber volumes using echocardiography on a large scale. The dataset includes 10,036 videos, each obtained during real clinical practice for diagnostic and medical decision-making purposes, involving different individuals. Originally intended as an external test set for our segmentation model, the dataset's apical-4-chamber view, specifically the heart angle in the videos, differs significantly from that in the CAMUS dataset. While our attempt to transfer our model to this external dataset was not successful, we include it here to document our exploration efforts.

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.

Comparing the Mask2Former variants, our ablation shows that larger architecture provides better segmentation results as expected. Also, we notice that implementing mask registration helps to save the losing performance in the Mask2Former experiments. Comparing the two investigated registration methods, the reg.b approach (applying deformation field to) provides result slightly better than the reg.b approach, but the difference is not significant.

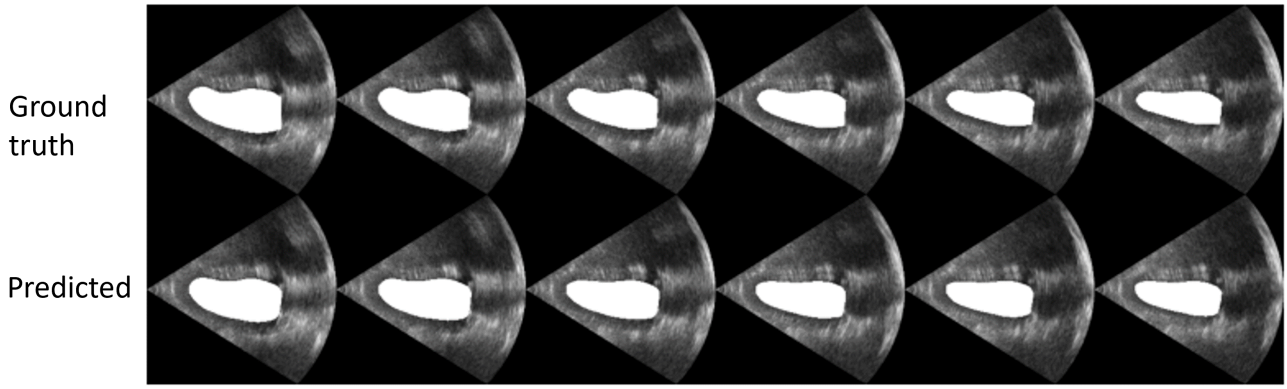

Figure 6. The plot displays the ground-truth masks alongside the predicted masks for the LV. It's important to note that the ground-truth mask, manually annotated by cardio-experts, features a sharp boundary between the LV and LA, which may not align with physiological natures. In contrast, owing to the temporal smoothness constraints, our predicted masks exhibit greater temporal smoothness and coherence across frames.

Method	DLabv3	Ours(DLabv3)	M2Fr (small)	M2Fr (small)	M2Fr (w/o reg.)	Ours (w/ reg.a)	Ours (w/ reg.b)
t-Dice	0.754	0.808	0.50	0.58	0.65	0.71	0.75

Table 1. Comparison between the vanilla DeepLabv3 (DLabv3), Mask2Former (M2Fr) and our proposed models with and without mask registration.

5.2. EF estimation

Unlike the success achieved in LV segmentation, our proposed model (DeepLabv3) resulted in a higher RMSE of 17.4, much worse than the previous best of 5.17. The Mask2Former-based model result in even worse 31.2 result, near random guesses. 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. We discuss the possible reason causing our failure in discussion.

6. Discussion

In this study, we introduce modules, such as temporal smoothness and mask registration, to enable two image models to understand temporal information from input videos. The evaluation of temporal segmentation indicates the effectiveness of these proposed methods. Compared to past results, we argue that the lower t-Dice scores are likely due to two factors. First, the previous reports were based on two key frames (ED/ES), not the temporal evaluations on all predicted masks, making it an unfair comparison. Second, due to time constraints, we primarily focus on validating the usefulness of the proposed method rather than optimizing the training process. For example, we only trained our models for 10 epochs, significantly less than the approximately 100 epochs used in previous studies. Judging from the effectiveness of the proposed method, we believe incorporating these modules into any state-of-the-art heart ultrasound models can help achieve a record-high score.

Unlike the success observed in LV segmentation, our model's performance in ejection fraction EF estimation is notably inadequate. We observe value instability during the process, where the predicted EF fails to properly converge to the supposed ground-truth, even after employing various modeling methods. Despite the challenges in predicting EF, we notice a phenomenon: adding EF as an auxiliary prediction task helps improve the performance of LV segmentation masks.

Contrary to our expectations, implementing Mask2Former did not lead to improvement. We attribute this to our unfamiliarity with transformer-based and instance segmentation models. The implementation-level unfamiliarity with the transformer-based model hinders effective debugging of the failure's cause. Notably, Mask2Former doesn't directly output masks but suggests sub-masks along with their corresponding attentions and queries. This complexity contrasts with the simpler implementation and debugging process of DeepLabv3.

Moreover, we encounter challenges in correctly establishing a one-to-one relation between predicted masks and their corresponding heart chambers. This issue can cause problems, especially when computing mask registration. For instance, we may condition the predicted mask for the LA on the predicted mask based on ED from LA and ES from the LV. This situation becomes more problematic when LA and LV resemble each other, a scenario not uncommon in the training data.

Looking ahead, I believe video understanding techniques can bring significant benefits to temporal segmentation in

heart ultrasound. Taking it a step further, there are additional heart ultrasound-related tasks that can only be effectively performed by video-based models. For instance, the concept of a "self-driving" heart catheter is a dream for cardiologists. Presently, these procedures rely on tactile methods, and leveraging video-based models could revolutionize and enhance these approaches.

In conclusion, despite the flaws in the failure analysis, we firmly believe that the proposed methods, particularly the mask registration approach, warrant further investigation and merit the pursuit of a research paper.

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. 3

[9] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015. 3

[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. 3

[11] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022. 3

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021. 3

[13] 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. 4