Supplementary material for "EchoDFKD: Data-Free Knowledge Distillation for Cardiac Ultrasound Segmentation using Synthetic Data"

Grégoire Petit[1,2,*], Nathan Palluau[*], Axel Bauer[2], Clemens Dlaska[1,2]
[1]Digital Cardiology Lab, Medical University of Innsbruck, A-6020 Innsbruck, Austria
[2]University Clinic of Internal Medicine III, Cardiology and Angiology,
Medical University of Innsbruck, A-6020 Innsbruck, Austria
g.petit360@gmail.com, nathan.palluau@gmail.com, clemens.dlaska@i-med.ac.at

Introduction

In this supplementary material, we provide:

- 1: Details regarding the varied sampling rates and different image qualities present in the dataset, including examples of corrupted clips.
- 2: Proof that the model outputs can be extremely close to the mean annotation of the annotator while being less noisy.
- 3: Extensive results due to the novelty of our approach and the lack of previous scores to compare against. The raw results can be found under `echoclip.csv`
- 4: When the model trained on synthetic data is evaluated on real data, most of the total error is concentrated on a small portion of the test examples.
- 5: Plots of $\log(scores)$ as a function of $\log(\text{Model\_size})$ for evaluation against human annotations and via EchoCLIP rewards.
- 6: An illustration of the poor performance in the very first frames, as noted in the main paper.
- 7: An extension of EchoDFKD to a multi-teacher setting where the right ventricle segmentation is also learned.
- 8: EchoDFKD inference on CAMUS dataset

1. Corrupted examples

As mentioned in Subsection 3.1. of the main paper, the EchoNet-Dynamic [5] dataset is very heterogeneous in terms of sampling rate (Figure 1), or image quality (Figure 3).

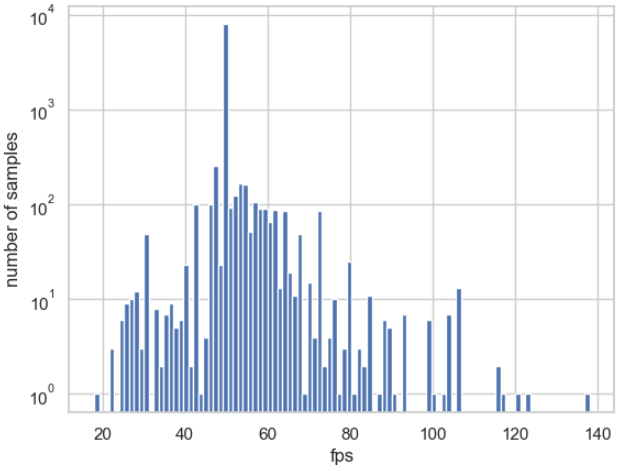


Figure 1. Distribution of sampling rates in the EchoNet-Dynamic [5] dataset.

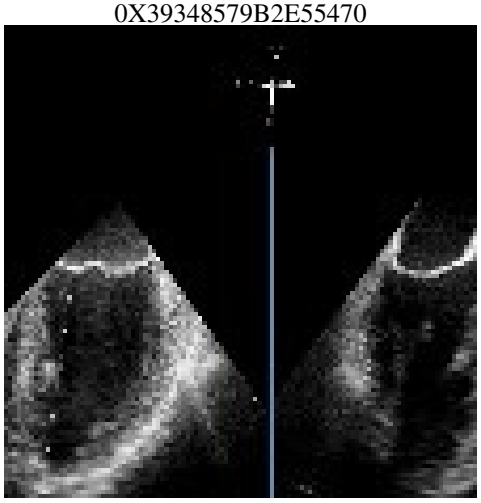
The most corrupted examples are 0X39348579B2E55470, 0X3693781992586497, and 0X790C871B162806D2, as displayed in Figure 2.

Additionally, we include in `corrupted.csv` different lists of corrupted examples for the following reasons:

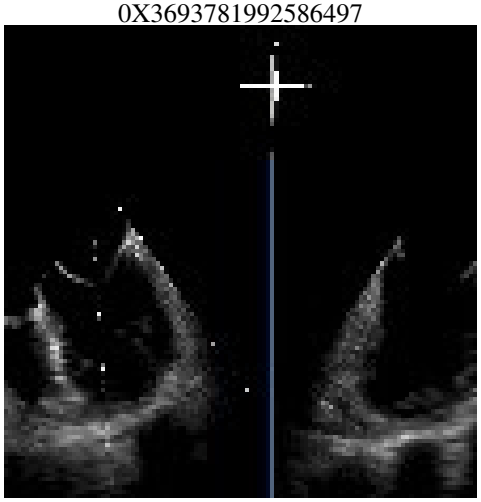
- Videos that are manifestly corrupted (see Figure 2).
- Labeling masks that are corrupted due to issues with [5]'s function fail to load labels properly for multi-labeled examples.
- Cases where the End-Systolic (ES) and End-Diastolic (ED) frames are too close together (differences close to one or even one in some examples).

[*]equal contribution

0X39348579B2E55470



0X3693781992586497



0X790C871B162806D2

Figure 2. 3 most corrupted examples.

2. Derivation of theoretical bounds for model scores in the function of intra-annotator scores

It seems that two slightly different definitions of intra-annotator standard deviation currently coexist in the literature. One considers the deviation between the values of a second annotation session and the first session, and the other considers the deviation between the values from one of the two sessions and a merged value from the two sessions (typically the mean). Here, since the RMSE reported in CAMUS is rather high compared to what we can observe from some models, we can infer that they used the first convention.

Consider two rounds of annotations, Z_1 and Z_2. We assume:

$$Z_1 = X_1 + Y$$
$$Z_2 = X_2 + Y$$

with X_1 and X_2 centered, i.i.d. (which is not very realistic but simplifies the derivations a lot), and Y being the latent truth (or, at least, the tendential value we would find with a lot of rounds).

The RMSE of the second round as an estimator of the first is :

$$\begin{aligned}\text{RMSE}(Z_2, Z_1) &= \sqrt{\mathbb{E}[(Z_2 - Z_1)^2]} \\ &= \sqrt{\mathbb{E}[(X_2 - X_1)^2]} \\ &= \sqrt{2\sigma_X^2}(\text{since } X_1 \text{ and } X_2 \text{ are independent}) \\ &= \sqrt{2} \cdot \sigma_X\end{aligned}$$

Now, the RMSE of a perfect model that would output Y, compared with a target obtained with a single annotation per example, is

$$\begin{aligned}\text{RMSE}(Y, Z) &= \sqrt{\mathbb{E}[(Z - Y)^2]} \\ &= \sqrt{\mathbb{E}[(X + Y - Y)^2]} \\ &= \sqrt{\mathbb{E}[X^2]} \\ &= \sigma_X\end{aligned}$$

We get:

$$\text{RMSE}(Z_2, Z_1) = \sqrt{2} \cdot \text{RMSE}(Z_2, Y)$$

Thus, the RMSE of Z_2 as an estimate of Z_1 is $\sqrt{2}$ times the RMSE of Z_2 with respect to Y.

CAMUS reports an intra-annotator std of 5.7. The theoretical lower bound of model performance is thus 4.03

On EchoNet-Dynamic, EchoCoTr reports an RMSE of 5.17

We can also look at the theoretical bound for correlation.

The correlation between Z and Y is :

$$\begin{aligned}\rho_{Z,Y} &= \frac{\text{Cov}(Z,Y)}{\sigma_Z\sigma_Y} \\ &= \frac{\text{Cov}(Y+X,Y)}{\sigma_Z\sigma_Y} \\ &= \frac{\sigma_Y^2}{\sigma_Z\sigma_Y} \\ &= \frac{\sigma_Y}{\sqrt{\sigma_Y^2+\sigma_X^2}}\end{aligned}$$

Next, the intra-annotator correlation, i.e. the correlation between Z_1 and Z_2 is :

$$\begin{aligned}\rho_{Z_1,Z_2} &= \frac{\text{Cov}(Z_1,Z_2)}{\sigma_{Z_1}\sigma_{Z_2}} \\ &= \frac{\text{Cov}(Y+X_1,Y+X_2)}{\sigma_{Z_1}\sigma_{Z_2}} \\ &= \frac{\sigma_Y^2}{\sigma_{Z_1}\sigma_{Z_2}} \\ &= \frac{\sigma_Y^2}{\sigma_{Z_1}^2} \\ &= \frac{\sigma_Y^2}{\sigma_Y^2+\sigma_{X_1}^2}\end{aligned}$$

Finally, we have :

$$\rho_{Z_1,Y} = \sqrt{\rho_{Z_1,Z_2}}$$

The best correlation coefficient that can be achieved between the model and the labeler, if only one annotation per example is available, is therefore $\sqrt{\rho_{Z_1,Z_2}}$.

CAMUS reports an intra-annotator correlation of 0.801. Thus, the theoretical upper bound is 0.895. [1], for instance, report a correlation of 0.78.

EchoCoTr doesn't provide a correlation. However, they report their R^2, which should be lower. After a linear regression between outputs and targets, the quadratic errors sum will be smaller (one would add two parameters that are allowed to fit the data used for evaluation); thus, R^2 of new outputs will be higher, and it will be equal to the correlation coefficient. Their squared correlation coefficient is higher than the reported R^2, making them close to the theoretical bound.

We also took an interest in the theoretical limit of a model's aFD score. This corresponds to the MAE of the selected frame number, and, through reasoning similar to the previous ones, it represents the average error of an annotator compared to a reference frame. We do not have access to this reference frame. Instead, we added an additional round of annotation by labeling over a thousand examples ourselves. This gives us an empirical distribution of $Z_1 - Z_2$, where Z_1 and Z_2 are the two rounds of labeling. By setting $Z_1 = Y + X_1$ and $Z_2 = Y + X_2$, with Y as the reference frame, and X_1 and X_2 iid, this reduces to $X_1 - X_2$, which distribution is obtained from that of X by convolution. Another practical assumption is to suppose that the distribution of X is the sum of a uniform distribution, which represents an annotator's abandonment when faced with a particularly degraded example (which happens very rarely, maybe once in several hundred examples, and results in discrepancies between two annotators that can reach up to fifty frames), and a symmetric distribution over a smaller support, which represents variations due to an annotator's lack of precision or a sequence of indistinct frames (resulting in discrepancies that rarely exceed nine or ten frames). We find that a Laplace distribution provides excellent log-likelihood after convolution by fitting different types of discrete distributions for the small support term. The expected distribution value obtained for $\|X\|_1$ is approximately 2.0 for the ES frame, and 2.4 for the ED frame.

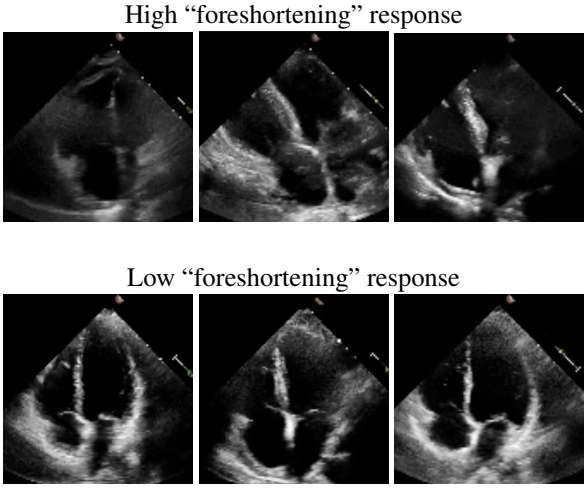


Figure 3. Example of samples depending on their EchoCLIP prompt response.

3. Extensive results

Our EchoCLIP [4] based raw results can be found under the `echoclip.csv` file of the `id636_supplementary.zip` file. Figure 3 represents the videos that activated the least and the most "foreshortening" prompts.

4. Portion of the dataset where the DFKD errors are located

When studying the proportion of EchoDFKD(DFKD) errors Our analysis observed that the total error is concen-

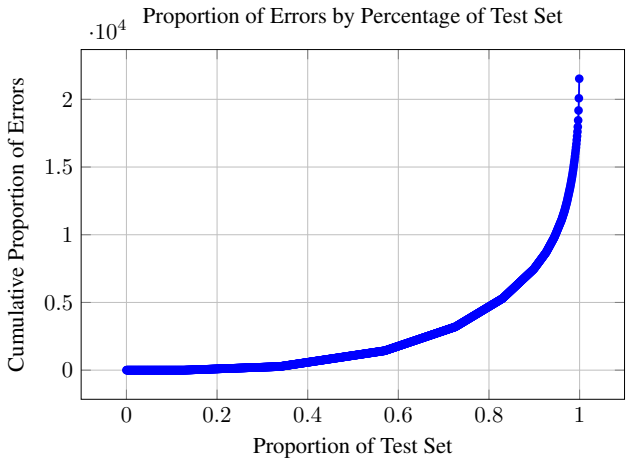


Figure 4. Proportion of squared errors as a function of test set percentage.

trated within a small portion of the test set. As depicted in Figure 4, the cumulative proportion of errors remains low for most of the test data, with a steep increase occurring in the final 15% of the dataset. This indicates that EchoDFKD performs well across most of the test set, and the errors are predominantly localized to a specific subset.

5. Scaling law

In Figures 5 and 6, we represented how 1-meanIoU, 1-Dice score and aFD_{ED}+aFD_{ES}scale with model weight. We observe that we reach a limit in performance around 1M parameters.

We obtain a log-slope of 0.15 for aFD versus human choice, 0.086 for aFD versus EchoCLIP, 0.16 for dice score, 0.11 for meanIoU. For comparison, in the regime with real data, aFD improves with a log-slope of 0.124 with human as a reference, 0.07 with EchoCLIP reference, 0.067 in meanIoU, and 0.088 in dice score. The slopes could provide insight into the dimension of the Riemannian manifold created by the model to handle its task. Still, it might be necessary to focus on precisely characterizing the boundary between the linear and saturation regimes and determining the theoretical limit with great precision to shift the logarithm instead of starting at 0 or 100%. We can at least observe that the slopes are steeper when training on synthetic data. Since the slopes tends to be inversely proportional to the dimensions of the surfaces [6], this is consistent with the idea that synthetic data tends to be less complex than real data.

6. First frames convergence

We mentioned in Section 4 of the main paper that one of the limitations of EchoDFKD was the few first frames for the model to converge to the solution, as depicted in Figure 7. In most cases, we can encounter that problem by taking the most significant connected component or prepadding the sequence to make the inferences converge.

7. Multi-teacher

Being able to accumulate multiple teachers to form a cohort opens up several possibilities, upon which the opportunity to refine target masks through ensemble learning and the ability to extend the single-task framework to multi-task learning [2].

Ensemble learning may require substantial computational resources for the model selection phase [3] and algorithms more sophisticated than simple averaging for combining the masks. Numerous variations of the standard STAPLE [7] algorithm have been adapted to address the specificities of a segmentation task.

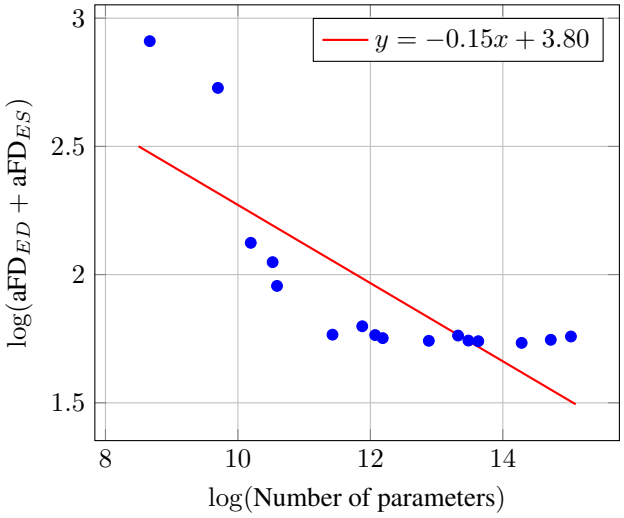
Here, we focus on the potential of multi-task learning. We trained our model to replicate the left ventricle masks of DeepLabV3 from Echonet Dynamics (as in the rest of the paper) and the right ventricle masks from another model, EchoGAN. While EchoGAN can also generate left ventricle masks, it is less precise than DeepLabV3 trained on Echonet Dynamics, having been trained on ten times fewer examples. We achieved performance comparable to the main experiment for left ventricle segmentation while simultaneously providing our model with a basic capability in right ventricle segmentation. For illustrative purposes, we show some outputs of the student model trained to segment the two ventricles in Figure 8.

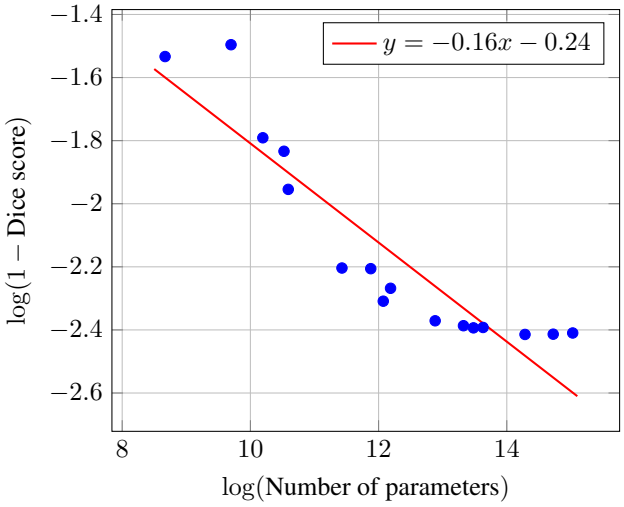
8. Inference on CAMUS dataset

The performance of EchoDFKD on CAMUS dataset is reported in Table 1. Despite its small size and short sequences, which penalize our model's warm-up requirements, compared to SimLvSeg Dice score (0.906), EchoDFKD still performs well (0.852) even though it's trained on synthetic data with far fewer parameters.

		B1	B2	B3	B4
meanIoU	l1	20.68%	68.89%	68.41%	73.37%
	l2	53.64%	66.44%	70.70%	75.08%
	l3	58.29%	64.46%	70.09%	72.69%
	l4	63.63%	49.81%	72.66%	73.56%
Dice score	l1	29.50%	80.29%	79.95%	83.65%
	l2	67.58%	78.21%	81.77%	85.21%
	l3	72.23%	76.20%	81.55%	85.03%
	l4	76.65%	62.08%	83.23%	84.17%

Table 1. Traditional performance metrics across EchoDFKD configurations, on the CAMUS dataset.





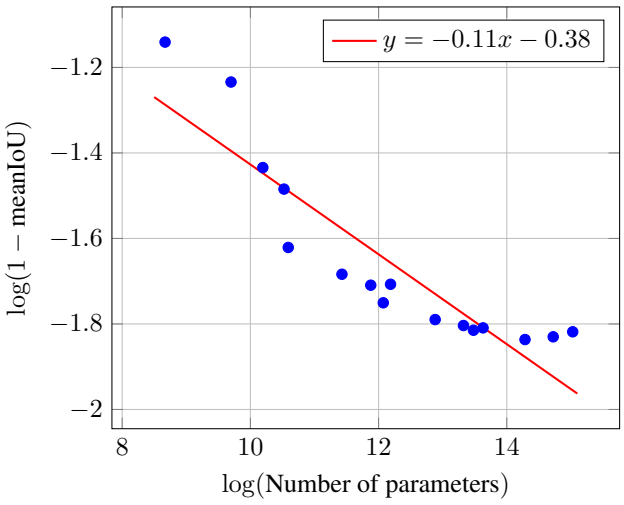


Figure 5. Scaling laws with humans as annotators.

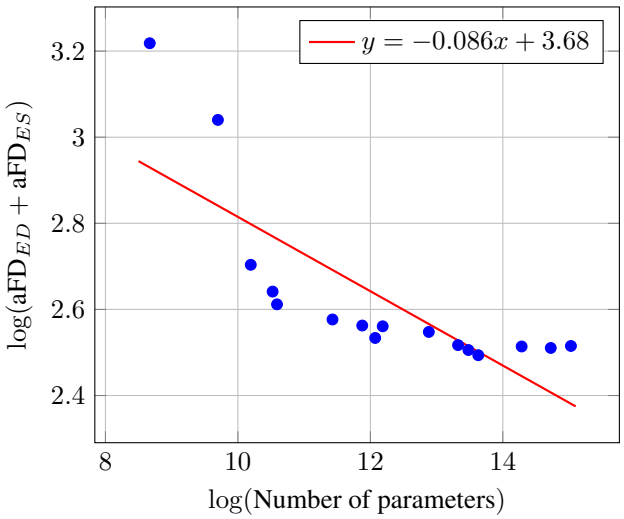


Figure 6. Scaling laws with EchoCLIP as annotator.

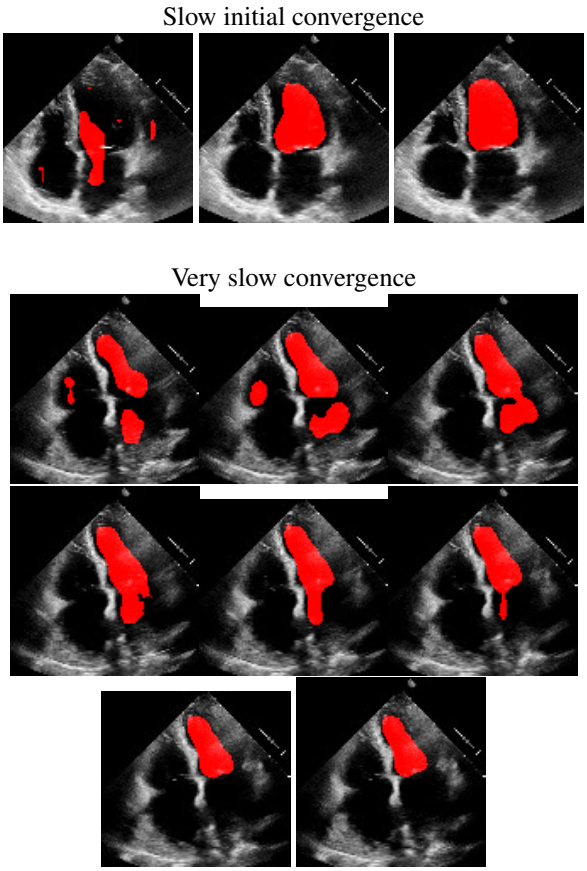


Figure 7. 2 examples of slow convergence.

(a) (b) (c)

Figure 8. EchoDFKD outputs when trained to segment the two ventricles.

References

[1] Samana Batool, Imtiaz Ahmad Taj, and Mubeen Ghafoor. Ejection fraction estimation from echocardiograms using optimal left ventricle feature extraction based on clinical methods. *Diagnostics*, 13(13):2155, 2023.

[2] Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.

[3] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning*, page 18, 2004.

[4] Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine*, pages 1–8, 2024.

[5] David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020.

[6] Utkarsh Sharma and Jared Kaplan. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23(9):1–34, 2022.

[7] Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–921, 2004.