Supplementary Materials for Generating Synthetic Labeled Data from Existing Anatomical Models: An Example with Echocardiography Segmentation

TABLE I
ABBREVIATIONS USED IN THE APPENDICES.

ED	End diastole
ES	End systole
A4C	Apical four chamber
A2C	Apical two chamber
LV	Left ventricle
LA	Left atrium
LV_{endo}	LV endocardium
LV_{epi}	LV epicardium

APPENDIX A
ADDITIONAL IMPLEMENTATION DETAILS

A. Original Anatomical Models

The original anatomical models used were a set of 19 models developed from 3D CT data by Rodero *et al.* [1]. These models will be made publicly available with the publication of [1] (currently under review). The models were developed to enable electromechanical simulations of a virtual patient cohort for cardiac resynchronization therapy and other treatments targeting heart failure. The model construction process mimicked that of Strocchi *et al.* [3] and full details on the construction process are available there. In brief, the models were constructed using a semi-automatic segmentation of 3D CT scans. The models include labels for all major cardiac anatomical regions, although valve anatomies are simplified to a 2mm thick plane since valve anatomies are not visible in the CT scans. Fig. 1 provides an example of the construction process. Note the level of detail in the models, which is a product of the high resolution in CT images.

B. New Anatomical Model Construction

A statistical shape model was constructed from the 19 original models. The shape model captures the distribution of cardiac shapes of the cohort of original meshes. To construct the shape model the meshes were rigidly aligned in 3D space, using the barycenters of the tricuspid valve, mitral valve and the lowest point of the interventricular septum. An arbitrarily chosen subject served as a reference for computation of rotation and translation matrices for all other cases. The alignment was performed to focus on quantifying the variability in the anatomies and attenuate the bias in construction of the mean shape.

Registration on the rigidly aligned surface meshes was performed using the Deformetrica software. The program allows for omitting the search for point-to-point correspondences and allowed for comparison based on the geometrical features where the interrelationships are non-parametric. The anatomical mean shape and the variability around it is computed from the surface meshes, represented with mathematical currents [4], [5]. In this process, every model can be obtained by applying subject

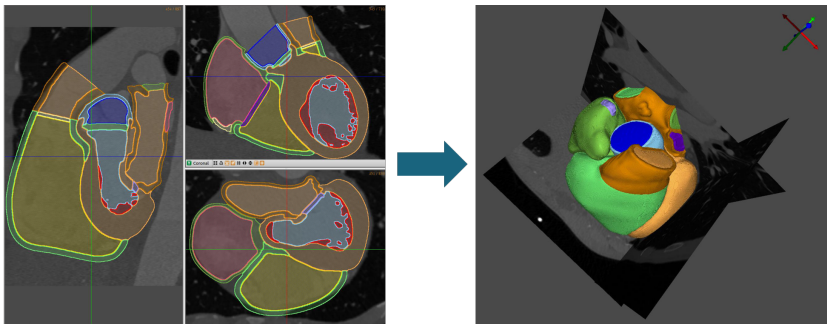

Fig. 1. Original model construction from Rodero *et al.* [1] using the Seg3D tool [2].

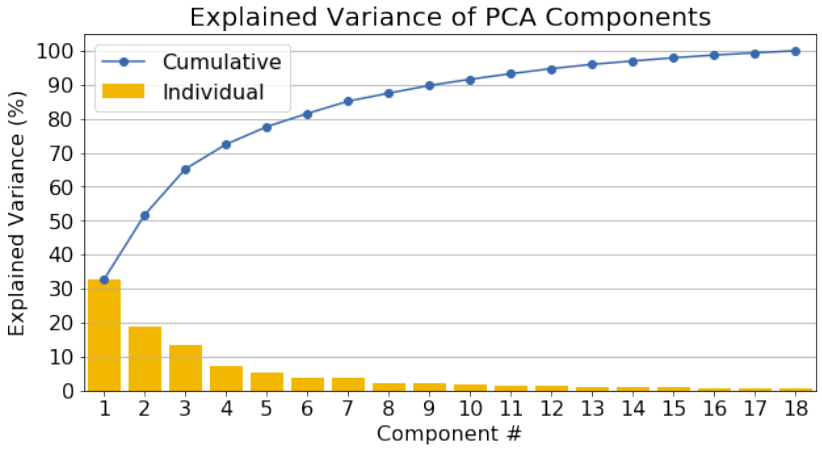

Fig. 2. The percentage of variance explained by each component after PCA.

specific transformation function to the template and adding residual information [6]. In this process, every model T_i from patient i can be obtained by applying subject specific transformation function ϕ_i to the template T^- and adding residual information as shown in Eq. (2) where $n \in \{0, 1\}$ and m is the shape mode [6].

$$H_j = (-1)^n \dot{\sum}_{k=0}^{8} m_k w_k^j \tag{1}$$

After the registration of the meshes, Principal Component Analysis (PCA) was applied to the deformation vectors characterised by functions ϕ_i to find the primary modes of shape variability within the population. The aim of PCA is to minimise the number of variables representing each sample while maximising the variance explained by these variables. The trade-off between model complexity and the amount of explained variability is controlled with the number of components chosen to represent the data set. The amount of variance explained by each component is shown in Fig. The 9 most prevalent shape modes computed with PCA captured over 90% of the variability and were chosen for the generative model.

The mesh k in the shape distribution of the original cohort can be approximated as a linear combination of the chosen PCA modes, weighted by certain weights w_k. We randomly sample each of these 9 modes within 2 standard deviations (square roots of the eigenvalues computed with PCA) to generate new synthetic meshes H_j as shown in

$$H_j = (-1)^n \dot{\sum}_{k=0}^{8} m_k w_k^j \tag{2}$$

where $n \in \{0, 1\}$ and m is the shape mode. Each of the synthetic models is still an anatomically plausible shape, but adds a heterogeneous example to our dataset. We repeat this procedure to generate 99 models in total. The generated surface meshes are transformed into volumetric meshes to allow for easy slice extraction.

The rigid transformation of the meshes, surface extraction, splitting the mesh into elements, labeling and merging was performed with Python programming language, using the VTK package [7]. Meshes were registered with Deformetrica software [8] and its atlas construction module. Transformation from surface to volumetric meshes was performed with gmsh [9].

C. 2D Slicing

Example two chamber and four chamber slices as well as their relationship are shown in Fig. 3. Valve center landmarks are defined as the center of mass of the respective tissue types while the apex is defined as the point in the LV myocardium which is farthest from the mitral valve center. The long axis is defined as the vector from the mitral valve center to the apex while the short axis is defined as the vector from mitral valve to tricuspid valve. These rotations were sampled from normal distributions with standard deviation of 16 and 9 degrees respectively. An example of extracting slices using varying rotations is shown in Fig. 3. Slicing was also implemented using the VTK package.

D. Pseudo Images

The generation of the pseudo images is application specific. It was performed based on an analysis of real A2C and A4C images. The first part of transforming the pseudo image is positioning the slice within the image. This consists of several steps:

1) An initial rough orientation of the slice. For A2C/A4C images this means rotating the slice such that the apex is at the top of the image and centering the LV within the image. Variations in the exact rotation and positioning are applied as an augmentation.

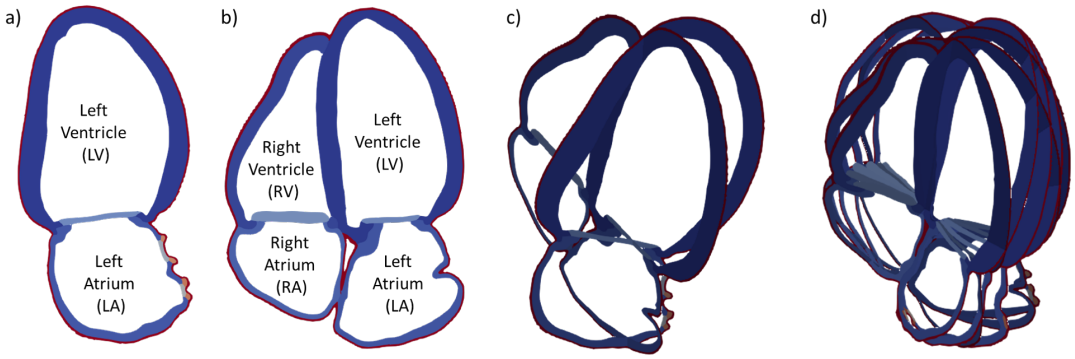

Fig. 3. Model slices as shown in Paraview [10]. a) Two chamber slice with labeled chambers in a model b) Apical four chamber slice with labeled chambers in a model c) Relationship between apical two chamber and apical four chamber slices in a model d) Example of how variable apical four chamber slices are extracted from the same model by varying rotation parameters.

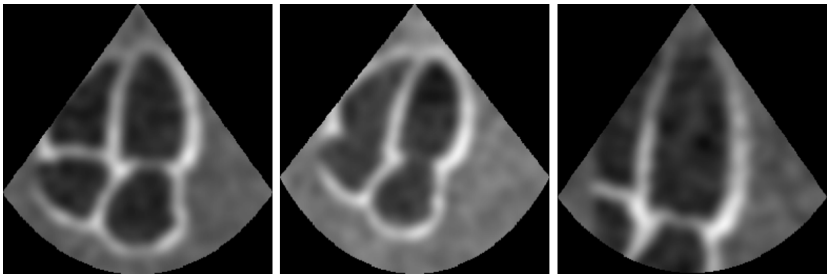

Fig. 4. Example of expanding the dataset using variable cropping parameters and affine transforms in the pseudo image construction step. All 3 pseudo images were constructed from the same slice. The first two images are full heart images while the last is LV-focused. The valve was removed from the middle image.

2) Following the real distribution, images were cropped to show either the entire heart or the left ventricle. Cropping was performed by cropping to $\pm$ 10% of the max/min locations of the relative attribute (pericardium for full heart images or LV myocardium for LV focused images). Fig. 4 shows an example of three different pseudo images created from the same slice where the first two are cropped to the pericardium and the last is cropped to the myocardium.
3) A2C/A4C images often are squeezed horizontally, particularly for full-heart images which much have a wider field of view to fit all chambers. When compressed to a 256 $\times$ 256 this has the effect of squeezing. Full heart images were squeezed by 30% and LV focused images were squeezed by 20%.
4) A cone of random width and depth was generated and used to mask the slice. The slice was shifted to ensure that the entire LV chamber was within the cone.

The second step of the pseudo transform is to modify the appearance to add noise and remove hard edges. The transformations applied to the slice consisted of several repetitions of a) adding /subtracting random uniform noise, b) multiplying by down-sampled random uniform noise, c) adding random shadowing or brightness in the form of 2D Gaussian kernels of varying size, d) shadowing the cone origin with the same method, and e) convolving with a Gaussian kernel to blur the image and remove hard edges. These transforms were done to add randomness to the pseudo images while also removing the hard edges (see Appendix F).

One of the primary problems for the segmentation network was correctly finding the mitral valve (see Appendix G). In order to try to increase the robustness of the network in finding the feature we randomly removed the model valve from 50% of the pseudo images (e.g. middle image of Fig. 4). We experimented to see if the CycleGAN would generate realistic valve features but it had no effect on preliminary results.

This step was performed individually for each dataset. The percentage of LV focused vs. whole heart images was set individually for each dataset based on qualitative analysis of a subset of 20 random images (60% for EchoNet and 50% for Camus/Site$_{\mathrm{A}}$) and image widths/depths were modified accordingly.

E. Transformation

Image flipping was disabled in the CycleGAN training since A4C and A2C images are always oriented in the same direction. Network architectures for the generator and discriminator are shown in Tables II and III respectively. The generator is a UNet [11] with 8 downsampling/upsampling layers using 2$\times$2 strided convolutions and instance normalization [12]. The discriminator is a PatchGAN with a 70$\times$70 pixel field of view. Both networks were implemented following [13]. The UNet contains 5.4M

TABLE II
NETWORK ARCHITECTURE FOR THE CYCLEGAN GENERATOR AND SEGMENTATION. THE NETWORK IS COMPOSED OF INPUT AND OUTPUT LAYERS AND 8 RECURSIVE UNET MODULE BLOCKS WHERE A UNET MODULE IS SHOWN BELOW. FILTER SIZES ARE DOUBLED DURING EVERY INITIALIZATION (STARTING AT 128) UNTIL REACHING A MAX VALUE OF 512.

Name	Type	Filters	Kernel	Stride	Activation	Normalization	Skip Connection
Input	Conv 2D	64	4×4	2×2			
UNet Module(F = 128)$^{\diamond}$	Conv 2D	F=128	4×4	2×2	Leaky ReLU	Instance 2D†	*
	UNet Module(min(F×2, 512))$^{\diamond}$						*
	Conv Transpose 2D	F=128	4×4	2×2	ReLU	Instance 2D	
Output	Conv Transpose 2D	1/2/4‡	4×4	2×2	ReLU		

Activation layers are processed before the main layer in the same row, normalization layers are processed after. $*$ indicates a skip connection between the outputs of the marked rows. $^{\diamond}$: the UNet Module is recursively included 8 times with filter size F doubling in each iteration until reaching 512. †: the Instance Norm 2D is not included in the innermost layer. ‡: Number of Output channels depends on the task.

TABLE III
NETWORK ARCHITECTURE FOR THE CYCLEGAN DISCRIMINATOR.

Name	Type	Filters	Kernel	Stride	Activation	Normalization
Input	Conv 2D	64	4×4	2×2		
L1	Conv 2D	128	4×4	2×2	Leaky ReLU	Instance 2D
L2	Conv 2D	256	4×4	2×2	Leaky ReLU	Instance 2D
L3	Conv 2D	512	4×4	2×2	Leaky ReLU	Instance 2D
Output	Conv 2D	128	4×4	2×2	Leaky ReLU	

Activation layers are processed before the main layer in the same row, normalization layers are processed after.

parameters and the PatchGAN contained 276k parameters. Other transformation parameters followed the defaults in [13]. Because the network struggles to generate the smooth cone outline of the ultrasound image, the label cone was used to mask the generated image during the final inference.

F. Segmentation

The segmentation network was also a UNet with 8 downsampling layers with the same architecture as Table II except the output channels were modified based on the task. We tested a smaller version of this architecture as well as the UNet1 and UNet2 architectures from [14] but found no increase in results in initial validation tests on the synthetic validation set. Preprocessing consisted of random cropping (256×256 crops from 286×286 images), gamma modifications (coefficient between 0.5 and 1.2), and mean normalization (specific to each dataset) when training. At test time all images were loaded as 256×256 and mean normalization was applied. The segmentation masks were post-processed to extract only the largest contiguous activation region and holes in the segmentation mask were filled. Learning rate was set to 1e-4 and decreased by a factor of 10 after 10 epochs. We used an Adam optimizer [15] and a batch size of 16. These parameters were set from previous experience with similar problems and a rough sweep while validating on the synthetic data. Within standard ranges these hyperparameters did not have a large impact on the performance of the segmentation network. We also experimented with several other standard segmentation loss functions including Dice loss and weighted cross entropy loss but found no increase in performance. Because the synthetic images are derived from models, they are more consistent than manual annotations and it is easier for the network to achieve high Dice scores on training and validation, which decreases the value of sweeping hyperparameters. All network training and inference was conducted in PyTorch 1.5.0 using an NVIDIA Titan GPU.

G. Datasets

A complete breakdown of dataset sizes by view and phase is shown in Table IV. 891 pseudo images were originally generated in the pipeline (19 models × 3 slices per model × 3 pseudo images per slice) but several were eliminated because the LV and LA regions were not adjacent (due to an improper cut plane). An individual pseudo dataset was generated for each corresponding synthetic dataset, but only one is shown in the table since the characteristics matched. The usage columns show how each dataset was used in the pipeline. The Camus, EchoNet, and Site_A real datasets were used to train the CycleGAN as well as for the label-free network selection described in the text. The synthetic datasets were used for supervised training of the segmentation networks. Finally, the Camus, EchoNet, Site_B, and Pathological real datasets were used to evaluate the networks and generate the results presented in this work.

TABLE IV
SIZES OF EACH DATASET INCLUDING TRAINING/VALIDATION/TEST SPLITS AND BREAKDOWNS BY VIEW AND PHASE.

Dataset	Annotator	Total	Split	Annotations			Size	Views		Phases		Usage		
				LV_{endo}	LV_{epi}	LA		A4C	A2C	ED	ES	Transform	Learn	Test
Camus [14]	[14]	1,600	train val test	✔	✔	✔	1280 160 160	656 80 82	624 80 78	640 80 80	640 80 80	○	○	●
EchoNet [16]	[16]	20,048	train val test	✔			14920 2576 2552	14920 2576 2552	0 0 0	7460 1288 1276	7460 1288 1276	○	○	●
$Site_A$	O_1	336	train val	✔	✔	✔	281 55	156 40	125 15	138 27	143 28	○	○	
$Site_B$	O_2	229	test	✔	✔	✔	229	149	80	115	114			●
Pathological	O_2	122	test	✔	✔	✔	122	122	0	61	61			●
Pseudo	Models	1754	train val	✔	✔	✔	1415 339	713 168	702 171	1415 339	0 0		●	
Synth Camus	Models	1754	train val	✔	✔	✔	1415 339	713 168	702 171	1415 339	0 0		●	
Synth EchoNet	Models	879	train val	✔	✔	✔	711 168	711 168	0 0	711 168	0 0		●	
Synth $Site_A$	Models	1754	train val	✔	✔	✔	1414 340	712 169	702 171	1414 340	0 0		●	

LV_{endo} = left ventricle endocardium, LV_{epi} = left ventricle epicardium, LA = left atrium, A4C = apical four chamber, A2C = apical two chamber, ED = end diastole, ES = end systole. ✔ indicates that the indicated annotation/view/phase is present in the dataset. For usage, ○ indicates that only the images were used (not labels) while ● indicates both images and labels were used. Transform, Segment, and Test correspond to b)-d) of Fig. 2 in the manuscript. Note that the usage markers includes only the proposed pipeline not comparison experiments.

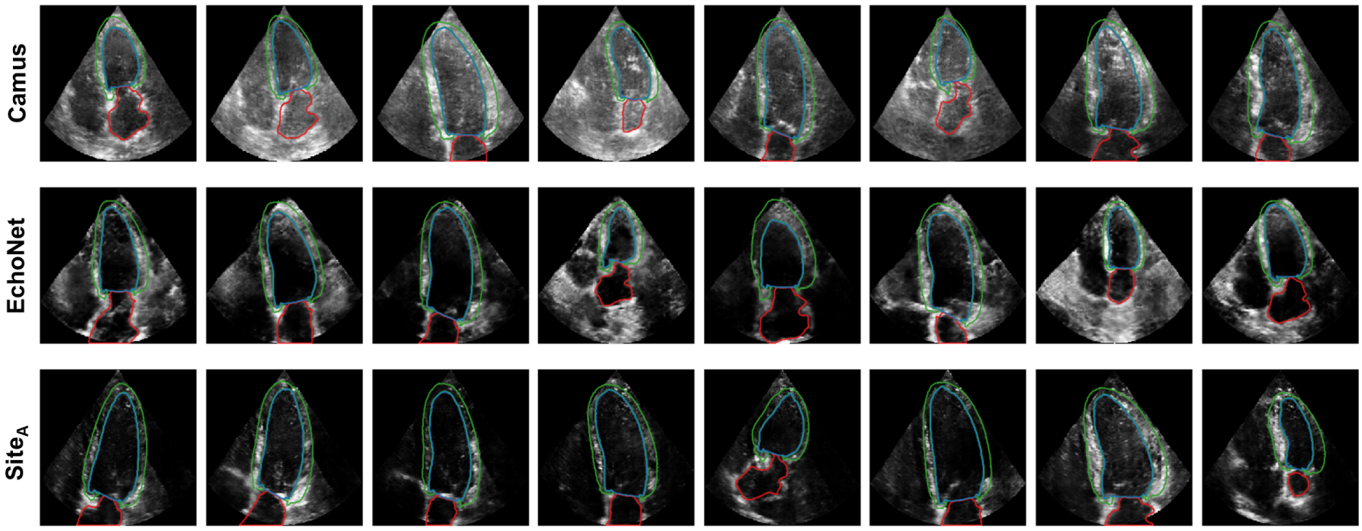

Fig. 5. 8 randomly selected synthetic images for each of the 3 synthetic datasets. Annotations are included, blue is LV_{endo}, green is LV_{epi}, and red is LA.

APPENDIX B
EXAMPLE SYNTHETIC IMAGES

Random example synthetic images from each of the 3 synthetic datasets are shown in Fig. 5. Note the difference in appearance of the different datasets which matches the difference seen in real images (see Appendix G for more example real images). While the images generally look realistic, the synthetic images can usually be identified because the noise pattern stays constant throughout the image. In real ultrasound the images are compressed at the top and stretched at the bottom because of the scan-conversion process, changing the noise characteristics throughout the image.

TABLE V
INTER-OBSERVER STUDY: MEDIAN RESULTS FROM A SECOND ANNOTATION ROUND ON 20 IMAGES FROM EACH DATASET. THE BOTTOM ROWS COMPARE ANNOTATIONS FROM O_1 AND O_2 FROM THE SECOND ANNOTATION ROUND.

Dataset	Version	Annotation Source		$\boldsymbol{D}$ (%) [$\boldsymbol{MAD}$]			$\boldsymbol{B}$ (%)			$\boldsymbol{S_p}$	$\boldsymbol{d_m}$
		Round 1	Round 2	LV_{endo}	LV_{epi}	LA	LV_{endo}	LV_{epi}	LA	LV_{endo}	
Camus	Real	[17]	O_1	78.7 [3.6]	90.6 [1.9]	85.5 [8.9]	42	17	3	0.83	10.0
			O_2	87.4 [4.4]	93.4 [1.7]	84.0 [6.9]	25	8	-5	0.83	5.7
	Synthetic	Models	O_1	83.8 [2.9]	88.6 [2.6]	78.0 [8.9]	30	16	26	0.80	6.7
			O_2	90.3 [2.3]	91.0 [2.3]	82.5 [9.3]	6	4	12	0.80	4.0
$Site_A$	Real	O_1	O_1	80.1 [5.1]	92.1 [2.2]	88.3 [3.5]	40	9	7	0.82	7.9
			O_2	88.9 [4.4]	93.3 [2.8]	90.2 [4.8]	21	8	-3	0.82	4.4
	Synthetic	Models	O_1	85.7 [4.1]	91.4 [2.5]	81.4 [4.9]	28	12	25	0.80	6.2
			O_2	91.1 [2.2]	92.5 [1.5]	83.3 [7.1]	11	4	15	0.80	3.9
EchoNet	Real	[16]	O_1	81.3 [5.4]	n/a	n/a	37	n/a	n/a	0.76	9.9
	Synthetic	Models	O_1	87.1 [2.3]	91.3 [1.7]	80.1 [5.6]	24	12	27	0.80	5.1
		Round 2	Round 2								
Camus & $Site_A$	Real	O_1	O_2	90.9 [3.2]	95.0 [2.0]	90.3 [4.5]	-17	-6	-10	0.85	4.3
	Synthetic	O_1	O_2	88.3 [4.1]	92.7 [3.4]	88.9 [3.2]	-21	-11	-10	0.84	4.9

LV_{endo} = left ventricle endocardium, LV_{epi} = left ventricle epicardium, LA = left atrium. $\boldsymbol{D}$ = Dice score, $\boldsymbol{B}$ = Bias percentage, $\boldsymbol{S_p}$ = simplicity, and $\boldsymbol{d_m}$ = mean average distance. $\boldsymbol{S_p}$ is listed for the dataset in the first column.

APPENDIX C
FULL INTER-OBSERVER RESULTS

Table V shows the full inter-observer results divided by dataset and observer. The bottom two rows show the comparison of the labels of O_1 and O_2 on the second round of labeling. The results show a significant difference between the first and second rounds of labeling by O_1. This is shown by the low Dice scores and high LV_{endo} biases in the intra-observer results on $Site_A$ as well as by the flip in biases between O_2 vs. O_1. In the first round O_2's LV_{endo} labels were significantly larger than O_1's (high positive bias) while in the second round they were significantly smaller. We estimate this difference comes from the differences in format of the second round of labeling (discussed in text) as well as the time difference (two months) in between the first and second rounds. It may also be due to other external factors.

APPENDIX D
ADDITIONAL SEGMENTATION RESULTS

A. Full results on A4C LV Segmentation Task

Quantitative results for testing the synthetically generated networks on the EchoNet, Camus, $Site_B$, and Pathological test sets for apical four chamber LV_{endo} segmentation in ED images only is shown in Table VI. The EchoNet results were already presented in the manuscript, but are shown again for comparison. Overall results show a similar pattern where the synthetic data approximately matches a network trained on a different real dataset. This varies by the test dataset with the synthetically trained network again performing well on the $Site_B$ data and Pathological data but worse on the Camus data.

B. Mean Results

The manuscript presents results using median and median absolute deviation because the results are not normally distributed, in which case the median is a better indicator of central tendency [18]. Mean and standard deviations for the results shown in Table I of the main manuscript are presented in Table VII for reference.

APPENDIX E
BREAKING DOWN OBSERVER BIAS

In this section we provide a more detailed analysis showing the most likely reason for LV_{endo} positioning differences (and corresponding differences in myocardial thickness) in the datasets is a bias between annotators rather than a difference in patient characteristics. Fig. 6 (O_1) and Fig. 7 (O_2) show violin plots for $\frac{LV_{endo}}{LV_{epi}}$ ratio on each of the different datasets O_1 and O_2 labeled in the inter-observer study. The results remain consistent across a given dataset. Fig. 8 shows the labeling ratio across different annotation rounds (i.e. between different annotators). In this case the ratio varies substantially between rounds. If the difference came from patient characteristics we would expect the ratios in Fig. 6 and Fig. 7 to vary between datasets. The converse indicates the primary difference in the ratio (and thus in LV_{endo} placement) instead comes from the labeler. These results also show that although O_1 was consistent within each labeling round there was a large bias between rounds as discussed above. This bias between labeling rounds was the reason the inter-observer results for O_1 were excluded from the main manuscript although results show the same pattern between real and synthetic data as O_2.

TABLE VI

TESTING ON REAL DATA: MEDIAN RESULTS FOR LV_{ENDO} SEGMENTATION NETWORKS TRAINED WITH SYNTHETIC AND REAL DATASETS AND EVALUATED ON REAL DATASETS OF APICAL FOUR CHAMBER IMAGES. RESULTS ARE ORDERED BY DICE SCORE. BOLD SHOWS THE BEST RESULT IN EACH SECTION.

Train	**Test**	$\boldsymbol{D}$ [$\boldsymbol{MAD}$]	$\boldsymbol{B}$ (%)	$\boldsymbol{S_p}$	$\boldsymbol{d_m}$
EchoNet	EchoNet	**94.0 [1.6]**	**0**	0.85	**2.7**
Camus		89.6 [2.6]	-2	**0.87**	4.8
$Site_A$		87.7 [3.8]	14	0.86	5.6
Synth EchoNet		87.1 [3.7]	15	0.85	5.9
Camus	Camus	**94.8 [1.2]**	**0**	0.85	**2.9**
EchoNet		93.0 [1.5]	-7	0.84	3.1
$Site_A$		89.6 [2.8]	14	**0.86**	4.8
Synth Camus		88.3 [2.6]	18	0.84	4.7
Synth $Site_A$	$Site_B$	**90.2 [2.4]**†	**-17**	0.84	**5.6**
EchoNet		88.3 [2.9]†,‡	-23	0.84	6.2
$Site_A$		87.9 [4.0]‡	-23	**0.85**	7.1
Camus		83.5 [4.3]	-32	0.85	8.8
EchoNet	Pathological	**92.0 [2.3]**	-7	0.86	**3.2**
Synth $Site_A$		90.7 [2.6]†	**-1**	0.85	4.1
$Site_A$		90.3 [3.5]†	**1**	0.86	4.2
Camus		89.5 [2.9]†	-10	**0.87**	4.6

$\dagger, \ddagger$: Rows in the same section marked with a matching symbol were not statistically different with a p-value < 0.05.

TABLE VII

MEAN METRICS COMPARING TRAINING WITH REAL DATASETS TO TRAINING WITH SYNTHETIC DATASETS. THE FIRST SECTION COMPARES INTER-OBSERVER RESULTS FOR O_2 ON REAL AND SYNTHETIC DATA. THE NEXT SECTION SHOWS NETWORKS TRAINED ON REAL AND SYNTHETIC DATA FOR LV_{ENDO} SEGMENTATION IN A4C ED IMAGES AND TESTED ON ECHONET. THE FINAL SECTION COMPARES NETWORKS TRAINED ON REAL AND SYNTHETIC DATA FOR ALL ANNOTATIONS/VIEWS/PHASES. ALL DICE RESULTS ARE STATISTICALLY DIFFERENT WITH A P-VALUE < 0.05 (COMPUTED WITH A WILCOXON SIGNED-RANK TEST). RESULTS ARE ORDERED BY DICE SCORE AND BOLD SHOWS THE BEST RESULT IN EACH SECTION.

Task	**Training Data** / *OR inter-observer comparison*	**Testing Data**	$\boldsymbol{D}$ (%) [$\boldsymbol{STD}$] LV_{endo}	LV_{epi}	LA	$\boldsymbol{B}$ (%) LV_{endo}	LV_{epi}	LA	$\boldsymbol{S_p}$	$\boldsymbol{d_m}$ LV_{endo}
Inter-Observer	*O_2 vs. Camus/$Site_A$*		86.9 [5.6]	**92.9 [2.8]**	**85.7 [9.6]**	24	7	**-4**	**0.85**	5.4
	O_2 vs. proposed pipeline		**90.2 [3.3]**	91.5 [3.0]	80.1 [14.5]	**8**	**2**	14	0.84	**4.1**
LV_{endo} in A4C ED	EchoNet (base)	EchoNet	**93.1 [3.5]**	n/a	n/a	**1**	n/a	n/a	0.85	**3.1**
	Camus		87.8 [7.4]			-2			**0.86**	6.2
	$Site_A$		86.1 [7.2]			15			**0.86**	6.2
	Synth EchoNet		85.6 [7.0]			16			0.84	6.7
All	Camus (base)	Camus	**92.5 [4.5]**	**95.2 [1.9]**	**88.3 [10.2]**	**2**	**-1**	**1**	0.83	**3.0**
	Site $_A$		87.5 [6.1]	91.0 [4.3]	79.8 [15.8]	10	-3	18	**0.84**	5.2
	Synth Camus		79.4 [9.9]	88.6 [5.4]	71.3 [23.7]	35	7	-4	0.82	8.4
	Synth $Site_A$	$Site_B$	**85.3 [8.8]**	88.1 [7.2]	73.7 [22.0]	**-4**	**3**	-11	0.82	**6.9**
	$Site_A$ (base)		85.1 [6.3]	**91.3 [4.6]**	**84.5 [12.5]**	-26	-9	16	**0.84**	7.9
	Camus		80.8 [12.0]	**91.3** [3.7]	83.0 [11.3]	-35	-9	**-4**	0.83	10
	Camus	Pathological	**88.2 [10.5]**	**91.1 [6.1]**	**88.6 [13.0]**	-8	**0**	**2**	**0.86**	**4.7**
	$Site_A$ (base)		86.5 [12.7]	89.6 [8.9]	83.8 [13.7]	**0**	9	14	0.85	5.3
	Synth $Site_A$		81.5 [17.8]	83.1 [12.7]	74.3 [23.5]	18	15	-20	0.82	6.5

LV_{endo} = left ventricle endocardium, LV_{epi} = left ventricle epicardium, LA = left atrium, A4C = apical four chamber, A2C = apical two chamber, ED = end diastole, ES = end systole. All refers to all annotations (LV_{endo}, LV_{epi}, and LA), views (A4C and A2C), and phases (ED and ES). $\boldsymbol{D}$ = Dice score, $\boldsymbol{MAD}$ = median absolute deviation, $\boldsymbol{B}$ = Bias percentage, $\boldsymbol{S_p}$ = simplicity, and $\boldsymbol{d_m}$ = mean average distance. For inter-observer $\boldsymbol{S_p}$ is listed for the second round annotations and $\boldsymbol{B}$ is calculated as $O_2 - $Original.

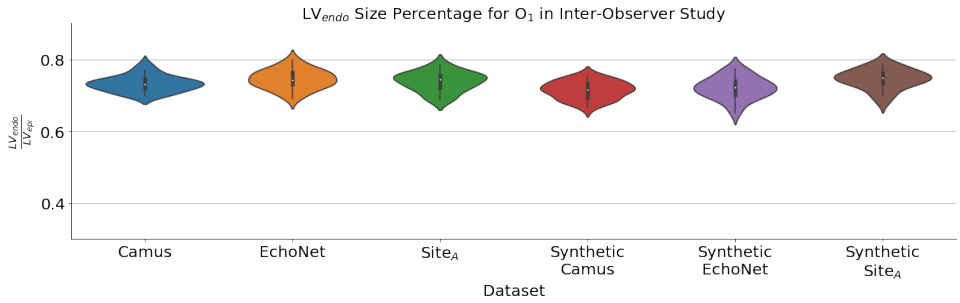

Fig. 6. LV_{endo} ratio for all datasets labeled by O_1 in the second round of labeling. The ratio stays consistent across the datasets.

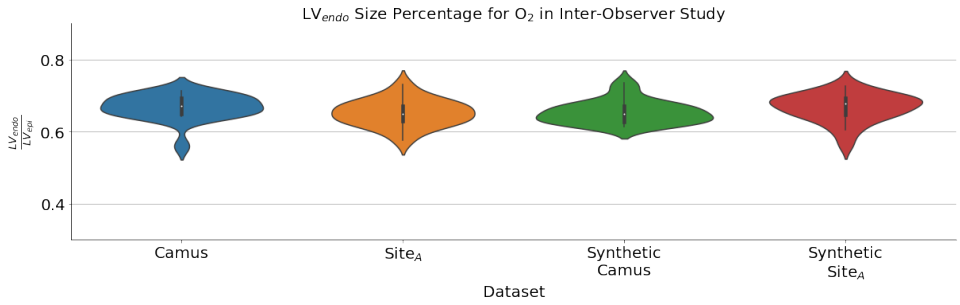

Fig. 7. LV$_{\text{endo}}$ ratio for all datasets labeled by O_2 in the second round of labeling. The ratio stays consistent across the datasets.

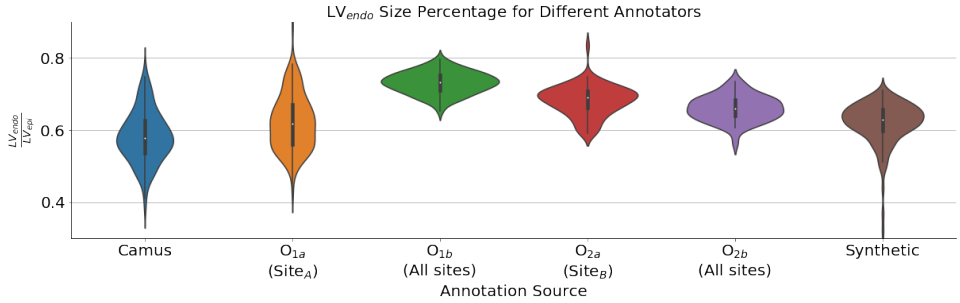

Fig. 8. LV$_{\text{endo}}$ ratio for each labeling round. The ratio varies substantially between rounds.

A qualitative evaluation of this phenomena is shown in Fig. 9. Example annotations are shown for each dataset in quantiles from 0% to 50% sorted by the $\frac{LV_{endo}}{LV_{epi}}$ ratio. The images show the much thicker myocardia in Camus and Site$_{\text{A}}$ even though the underlying tissue does not appear significantly different in many cases. Pseudo images are presented for the Synthetic dataset.

APPENDIX F
EFFECT OF PSEUDO TRANSFORM

To evaluate the effect of the pseudo transforms (Appendix A-D) several experiments were conducted. Segmentation networks were trained using images from the model after various versions of the pseudo and CycleGAN transformations as shown in Fig. 10. Networks were trained with images directly from the model (”slice images”), images with the pseudo appearance transforms (”pseudo images”), slice images transformed with a CycleGAN (”synthetic slice images”), and pseudo images transformed with a CycleGAN (”synthetic images”). Qualitatively, the synthetic slice images do not appear as realistic as the synthetic images. Many of the hard edges are maintained from the slice image making it easy to tell that the images are fake.

Segmentation results are shown in Table VIII. Results indicate that the synthetic images achieve the best results as shown in the manuscript. The slice image dataset achieves poor results as the network trained on this dataset was not able to translate learned features across the large texture differences. The networks trained on the synthetic slice dataset and the pseudo dataset were approximately the same and slightly worse than the network trained on the synthetic images.

TABLE VIII

COMPARING TRANSFORMATIONS: MEDIAN RESULTS FOR LV$_{\text{ENDO}}$ SEGMENTATION NETWORKS TRAINED ON THE DATASET VARIATIONS SHOWN IN FIG. 10. RESULTS ARE ORDERED BY DICE SCORE ($\boldsymbol{D}$). ALL DICE RESULTS EXCEPT THOSE MARKED (†) ARE STATISTICALLY DIFFERENT WITH A P-VALUE $<$ 0.05 (COMPUTED WITH A WILCOXON SIGNED-RANK TEST). BOLD SHOWS THE BEST RESULT FOR EACH METRIC.

Train	**Test**	$\boldsymbol{D}$ (%) [$\boldsymbol{MAD}$]	$\boldsymbol{B}$ (%)	$\boldsymbol{S_p}$	$\boldsymbol{d_m}$
Synthetic EchoNet	EchoNet	**87.1 [3.7]**	**15**	**0.85**	**5.9**
Synthetic Slice EchoNet		84.6 [4.9]†	**15**	0.84	6.8
Pseudo EchoNet		84.1 [4.5]†	18	0.84	6.7
Slice EchoNet		65.3 [11.8]	**15**	0.80	12.5

†: Not statistically different with a P-value $<$ 0.05.

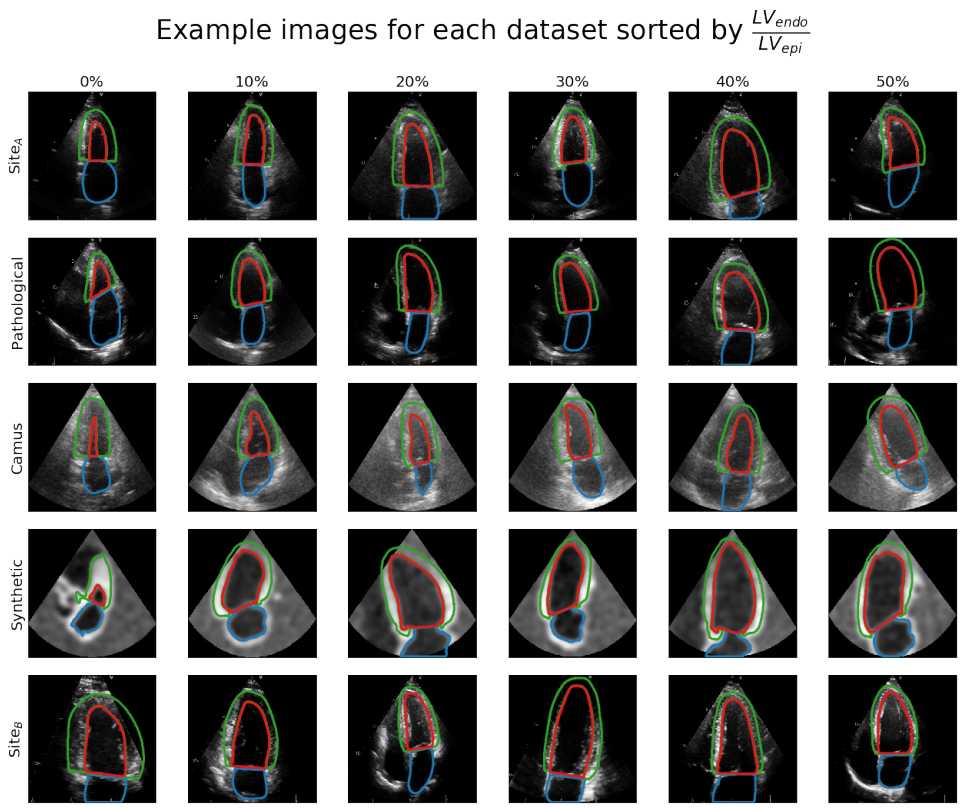

Fig. 9. Example annotation results for each dataset sorted by $\frac{LV_{endo}}{LV_{epi}}$ ratio within each dataset. Each column shows a quantile of the $\frac{LV_{endo}}{LV_{epi}}$ ratio from 0% to 50%. The 0% quantile represents the image with the lowest $\frac{LV_{endo}}{LV_{epi}}$ ratio (largest myocardium). As shown, the Camus dataset typically has much larger myocardia.

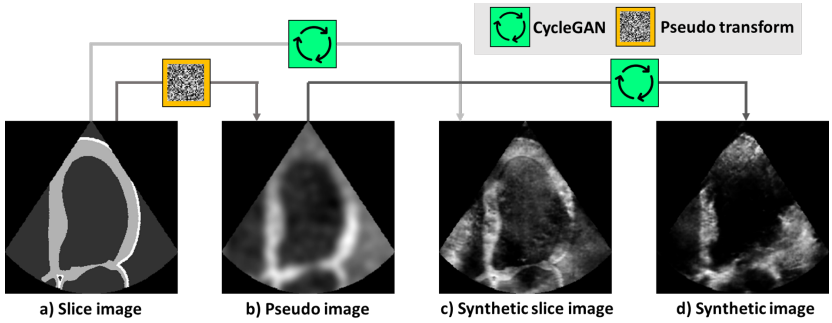

Fig. 10. **Four different image types were tested to determine the effects of the pseudo and CycleGAN transforms:** a) Slice images which are slices extracted from the model with only a cone mask applied, b) pseudo images which have undergone the pseudo transformations described in the manuscript, c) synthetic slice images which are the slice images passed through a CycleGAN, and d) synthetic images which are the pseudo images passed through the CycleGAN.

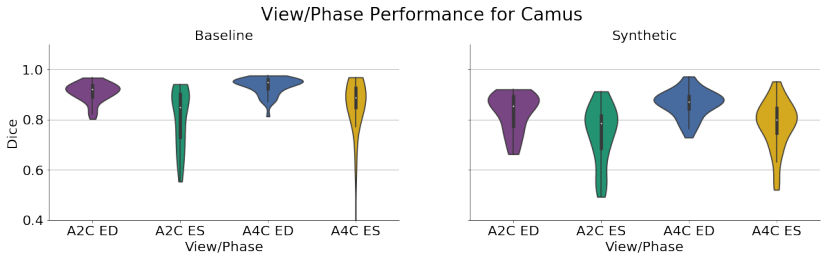

Fig. 11. **View/phase breakdown for Camus:** Performance split by echo views and phases for the Camus test set. The network trained on synthetic data shows much worse results for the end systole (ES) phase than end diastole (ED). However, this is also matched by the baseline network indicating that the ES images are more difficult in general for this dataset. This finding matches [14]. A2C = apical two chamber, A4C = apical four chamber, ED = end diastole, ES = end systole.

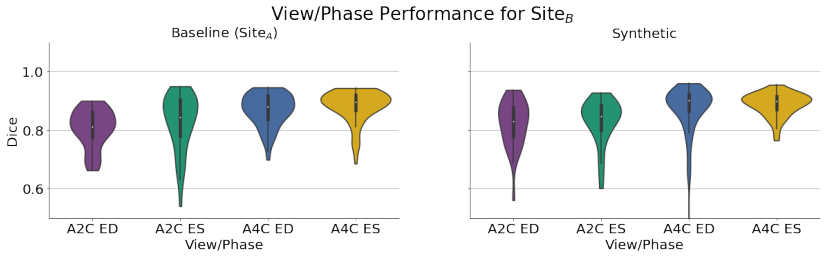

Fig. 12. **View/phase breakdown for Site$_B$:** Performance split by echo views and phases for the Site$_B$ test set. In this case the results are approximately consistent across views and phases for both the baseline and synthetic experiments. A2C = apical two chamber, A4C = apical four chamber, ED = end diastole, ES = end systole.

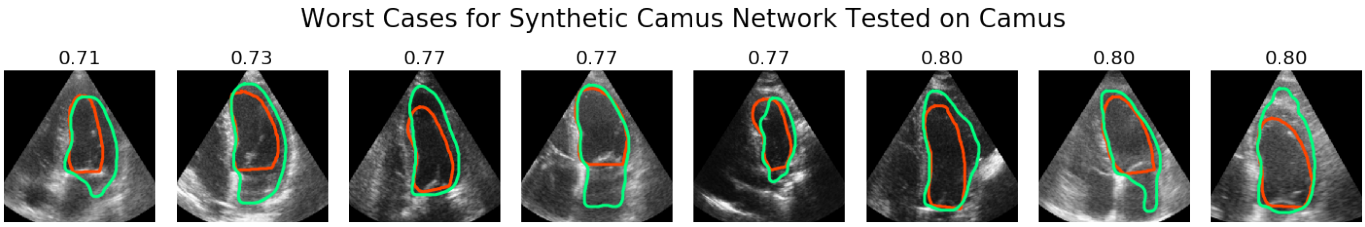

Fig. 13. The 8 worst images (sorted by Dice score) for the network trained on synthetic Camus data and tested on real Camus data. Titles show Dice score. Red shows the label while green shows the network output.

Appendix G
Failure Analysis

A. Breakdown by phase and view

Fig. 11 and Fig. 12 show the breakdown in performance for different echo views and phases for the Camus and Site$_B$ test respectively. Results are varied between the two sets with Camus showing a clear difference in performance between ED and ES for both the synthetic model and the baseline model. While we initially hypothesized the difference on synthetic data was because the model did not include ES images, the equivalent result on real images indicates the difference is likely an implicit bias with ES images in the dataset. Site$_B$ on the other hand shows approximately equivalent performance across the phases and views for both although ES is slightly higher than ED for the baselinereal-data model. This discrepancy could come from several sources, with the most likely being the selection criteria for determining the phase.

In both cases the synthetic four chamber images perform slightly better than the synthetic two chamber images. This is likely because the extraction process was optimized for four chamber images, meaning those images are of better quality.

B. Worst Cases for Segmentation

We focused this section on LV$_{endo}$ segmentation only to simplify the analysis, but the reasons for failure are similar when extending to additional annotations. Fig. 13, Fig. 14, Fig. 15, and Fig. 16 show the worst case images for each of the networks trained on the synthetic versions of Camus, Site$_A$, and EchoNet and tested on the corresponding real datasets. We analyzed these images to identify the primary sources of failure for the network:

1) In many cases the network is unable to properly detect the mitral valve and includes part of the LA in the LV. Because the valve is not visible in CT images it is modeled as a flat slab in the models. This means the synthetic images do not include the variety of valves that are included in the real images. Including accurate valve representations is a target for future work.

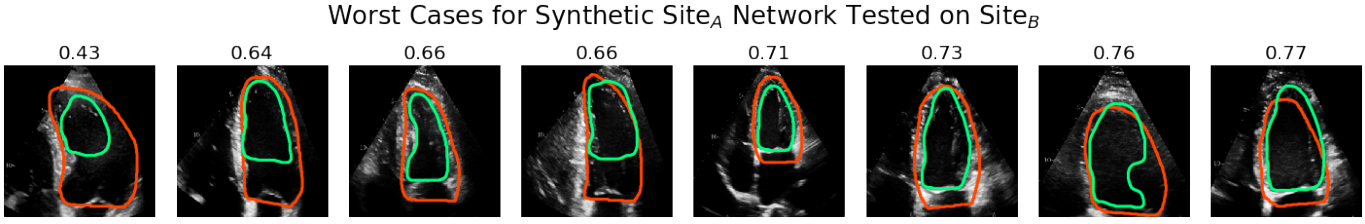

Fig. 14. The 8 worst images (sorted by Dice score) for the network trained on synthetic Site$_A$ data and tested on Site$_B$. Titles show Dice score. Red shows the label while green shows the network output.

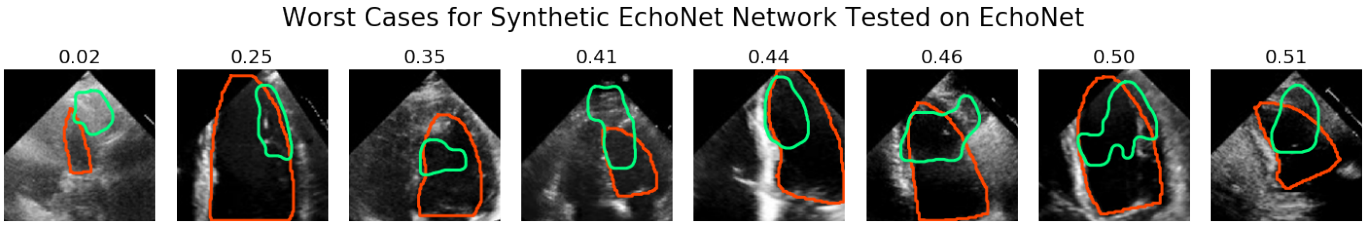

Fig. 15. The 8 worst images (sorted by Dice score) for the network trained on synthetic EchoNet data and tested on real EchoNet data. Titles show Dice score. Red shows the label while green shows the network output.

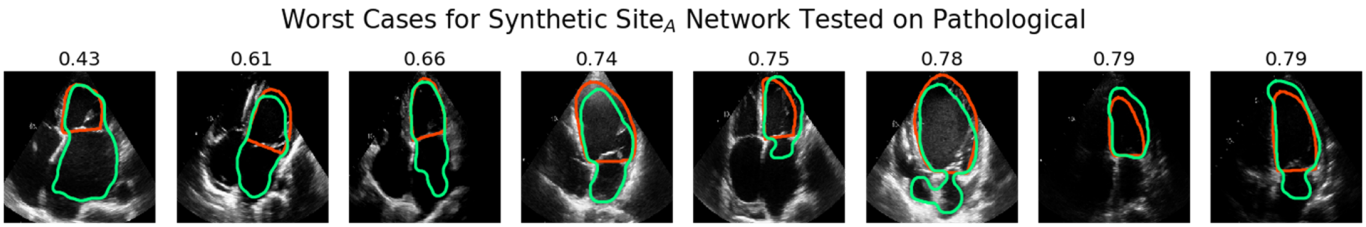

Fig. 16. The 8 worst images (sorted by Dice score) for the network trained on synthetic Site_A data and tested on the Pathological dataset. Titles show Dice score. Red shows the label while green shows the network output.

2) The network fails for some LV-focused images where the mitral valve is positioned at the bottom edge of the image. This is likely due to a difference in the pseudo dataset construction as all pseudo images were cropped such that there was at least a 10% border between the mitral valve and the bottom edge of the image. In other cases the walls were cropped out by the cone which was also avoided in the pseudo images.
3) In some cases the walls were dropped out of the image due to the high gain settings. With only a single frame it can be nearly impossible to detect the wall so giving additional frames to the segmentation network would be the best solution to address this. An approach such as random erasing augmentation [19] may also help resolve these cases.
4) In addition, the network appeared to struggle in cases where the image was wider or where the LV was slightly rotated. These factors of variability could be included in the pseudo image generation process.
5) The network also struggled in some cases where the image was wider or where the LV was slightly rotated.
6) In the pathological dataset the network struggled in some cases with high left atrial dilation where the left atrium is significantly larger than the left ventricle.

These items played a varying role in failure cases for the different datasets and illustrate the challenges of deploying an algorithm into the wild.

While we have presented the worst cases for the synthetic segmentation networks above, in general the networks perform very well. Fig. 17 shows a result from each quantile of the Dice score for a network trained on the baseline and synthetic versions of each dataset. As shown, the majority of the images have good segmentation results. For example, although there were many bad failures for EchoNet (Fig. 15), the test set was much larger in this case, providing more opportunities for failure. In general, the network performed well.

Appendix H
Pre-training With Synthetic Data

The primary goal of the proposed pipeline is to generate artificial data. However, real data can also be used to fine-tune the results from the synthetic data model. This may be especially useful in cases where only a small number of real images are available. Fig. 18 shows results of fine-tuning a model pretrained with synthetic data on varying amounts of real images compared to training from scratch with real images. Results show that for a small number of images, pre-training with synthetic data yields significantly higher Dice scores. This is particularly true for the LA which the model initially struggles to segment accurately (matching the trend reported in [17]). As more real images were added there are diminishing returns from pre-training with synthetic data, but in almost all cases the pre-trained model has better performance. Note that in the case of LV_endo segmentation trained on Site_A, the model trained on synthetic data initially outperformed the model trained on real data when tested on Site_B. In this case, as more real data images are added there is a small decrease in performance in LV_endo Dice scores. Fig. 19 shows the performance improvements from pretraining on the synthetic data on the Camus and Site_A datasets for each segmentation target.

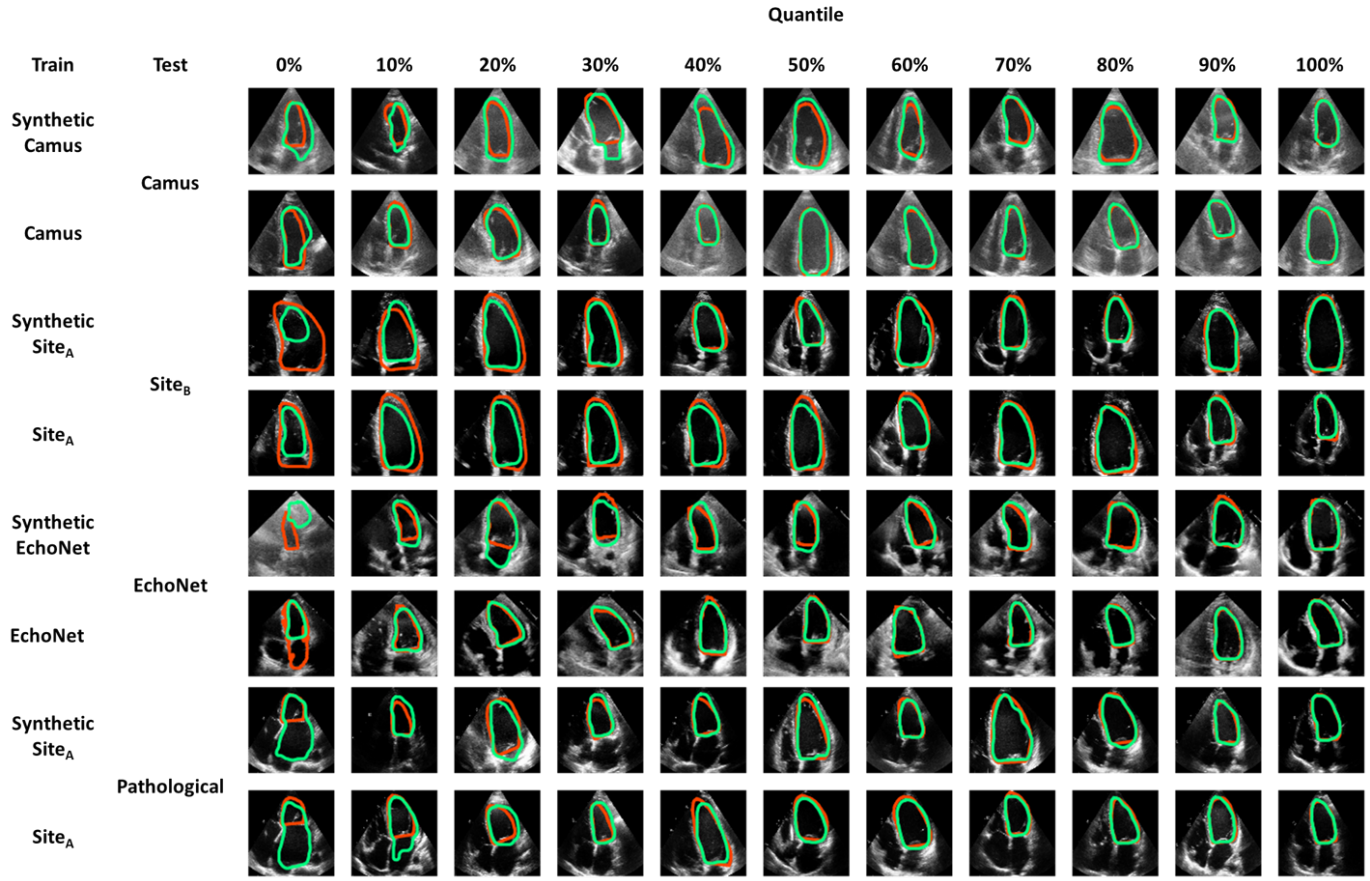

Fig. 17. An example image from each quantile of the Dice results for LV_{endo} segmentation in apical four chamber images. Each row shows results for a network trained on a different dataset. Red shows the label while green shows the network output.

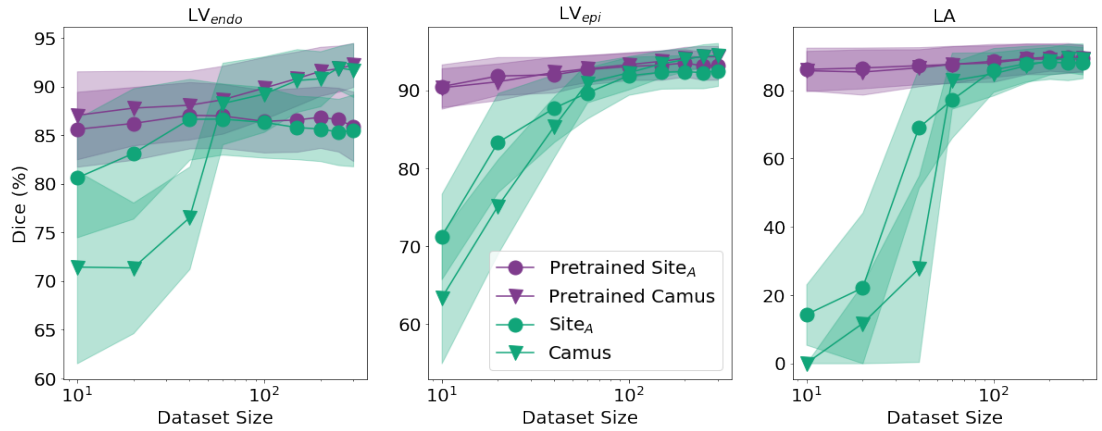

Fig. 18. Pretraining with synthetic data can improve results if only limited data is available. Median LV_{endo}, LV_{epi}, and LA Dice scores for networks pretrained on synthetic data vs. networks trained from scratch. Error bars show median absolute deviation. The networks trained on Camus were tested on Camus and the networks trained on $Site_A$ were tested on $Site_B$. X-axis is log scale.

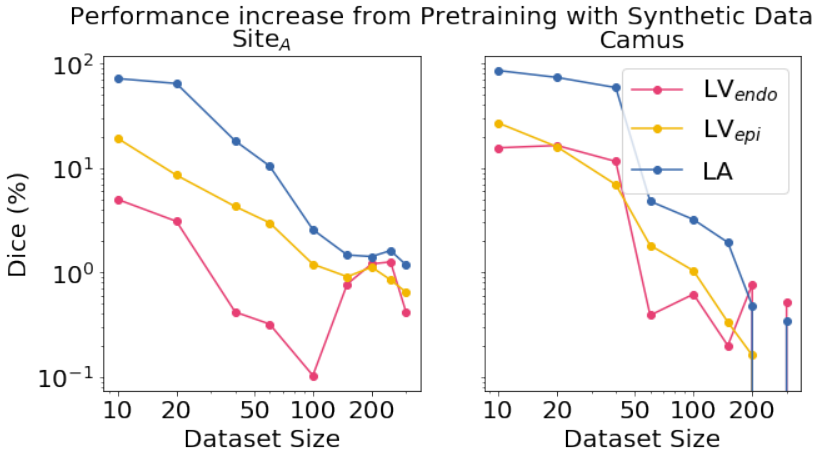

Fig. 19. The performance improvement from pretraining with synthetic data for Site$_A$ and Camus. Both axes are shown on a log scale.

REFERENCES

[1] C. Rodero, M. Strocchi, M. Marciniak, J. Whitaker, D. O. Neill, K. Gillette, C. Augustin, G. Plank, E. Vigmond, P. Lamata, and S. A. Niederer, "Anatomical changes influences mechanics, electrophysiologyand haemodynamics in a complementary and localised wayin the healthy adult human heart," *PLOS Comput. Biol. (under Rev.)*.

[2] CIBC, "Seg3D: Volumetric Image Segmentation and Visualization," 2016. [Online]. Available: http://www.sci.utah.edu/software/seg3d.html

[3] M. Strocchi, C. M. Augustin, M. A. Gsell, E. Karabelas, A. Neic, K. Gillette, O. Razeghi, A. J. Prassl, E. J. Vigmond, E. J. Vigmond, J. M. Behar, J. M. Behar, J. Gould, J. Gould, B. Sidhu, B. Sidhu, C. A. Rinaldi, C. A. Rinaldi, M. J. Bishop, G. Plank, and S. A. Niederer, "A publicly available virtual cohort of fourchamber heart meshes for cardiac electromechanics simulations," *PLoS One*, vol. 15, no. 6 June, pp. 1–26, 2020. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0235145

[4] M. Vaillant and J. Glaunès, "Surface Matching via Currents," in *Inf. Process. Med. Imaging*, G. E. Christensen and M. Sonka, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 381–392.

[5] T. Mansi, I. Voigt, B. Leonardi, X. Pennec, S. Durrleman, M. Sermesant, H. Delingette, A. M. Taylor, Y. Boudjemline, G. Pongiglione, and Others, "A statistical model for quantification and prediction of cardiac remodelling: Application to tetralogy of fallot," *IEEE Trans. Med. Imaging*, vol. 30, no. 9, pp. 1605–1616, 2011.

[6] S. Durrleman, X. Pennec, A. Trouvé, and N. Ayache, "Statistical models of sets of curves and surfaces based on currents," *Med. Image Anal.*, vol. 13, no. 5, pp. 793–808, 2009.

[7] W. Schroeder, K. Martin, and B. Lorensen, *The Visualization Toolkit*, 4th ed. Kitware, 2006.

[8] S. Durrleman, M. Prastawa, N. Charon, J. R. Korenberg, S. Joshi, G. Gerig, and A. Trouvé, "Morphometry of anatomical shape complexes with dense deformations and sparse parameters," *Neuroimage*, vol. 101, pp. 35–49, 2014. [Online]. Available: www.deformetrica.org.

[9] C. Geuzaine and J.-F. Remacle, "Gmsh: A 3-D finite element mesh generator with built-in pre-and post-processing facilities," *Int. J. Numer. Methods Eng.*, vol. 79, no. 11, pp. 1309–1331, 2009.

[10] J. Ahrens, B. Geveci, and C. Law, *ParaView: An End-User Tool for Large Data Visualization,*. Elsevier, 2005.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Int. Conf. Med. image Comput. Comput. Interv.*, 2015.

[12] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance Normalization: The Missing Ingredient for Fast Stylization," no. 2016, 2016. [Online]. Available: http://arxiv.org/abs/1607.08022

[13] F. Jay, J.-P. Renou, O. Voinnet, and L. Navarro, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks Jun-Yan," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 183–202.

[14] S. Leclerc, E. Smistad, A. Ostvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, C. Lartizien, L. Lovstakken, and O. Bernard, "Deep Learning Segmentation in 2D echocardiography using the CAMUS dataset : Automatic Assessment of the Anatomical Shape Validity," in *Int. Conf. Med. Imaging with Deep Learn. – Ext. Abstr. Track*, 2019.

[15] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–15, 2015.

[16] D. Ouyang, B. He, A. Ghorbani, C. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, J. Euan A Ashley, and a. Y. Zou, "Interpretable AI for beat-to-beat cardiac function assessment," *Nature*, 2020. [Online]. Available: https://doi.org/10.1101/19012419

[17] S. Leclerc, E. Smistad, J. Pedrosa, A. Ostvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P. M. Jodoin, T. Grenier, C. Lartizien, J. Dhooge, L. Lovstakken, and O. Bernard, "Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography," *IEEE Trans. Med. Imaging*, vol. 38, no. 9, pp. 2198–2210, 2019.

[18] M. Pagano and K. Gauvreau, *Principles of Biostatistics*, 2nd ed. CRC Press, 2018.

[19] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random Erasing Data Augmentation," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, pp. 13 001–13 008, 2020.