UNSUPERVISED ECHOCARDIOGRAM VIEW DETECTION VIA AUTOENCODER-BASED REPRESENTATION LEARNING

Andrea Treviño Gavito
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208
andrea.tg@u.northwestern.edu

Diego Klabjan
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208
d-klabjan@northwestern.edu

Sanjiv J. Shah
Feinberg Cardiovascular Research Institute
Northwestern University
Chicago, IL 60611
sanjiv.shah@northwestern.edu

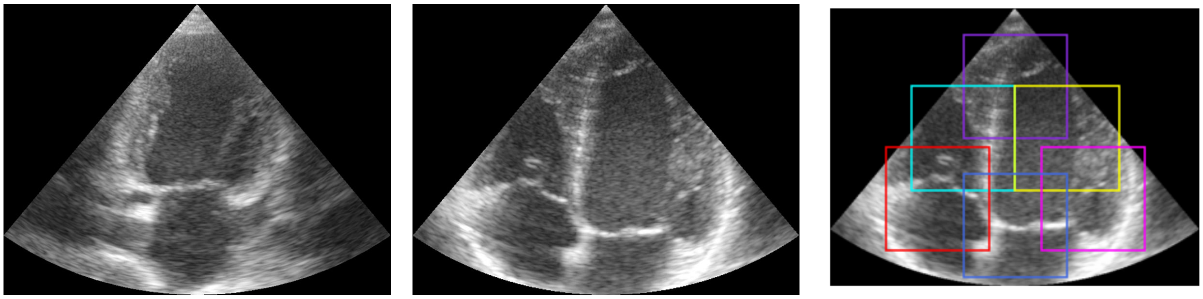
ABSTRACT

Echocardiograms serve as pivotal resources for clinicians in diagnosing cardiac conditions, offering non-invasive insights into a heart's structure and function. When echocardiographic studies are conducted, no standardized labeling of the acquired views is performed. Employing machine learning algorithms for automated echocardiogram view detection has emerged as a promising solution to enhance efficiency in echocardiogram use for diagnosis. However, existing approaches predominantly rely on supervised learning, necessitating labor-intensive expert labeling. In this paper, we introduce a fully unsupervised echocardiographic view detection framework that leverages convolutional autoencoders to obtain lower dimensional representations and the K-means algorithm for clustering them into view-related groups. Our approach focuses on discriminative patches from echocardiographic frames. Additionally, we propose a trainable inverse average layer to optimize decoding of average operations. By integrating both public and proprietary datasets, we obtain a marked improvement in model performance when compared to utilizing a proprietary dataset alone. Our experiments show boosts of 15.5% in accuracy and 9.0% in the F-1 score for frame-based clustering, and 25.9% in accuracy and 19.8% in the F-1 score for view-based clustering. Our research highlights the potential of unsupervised learning methodologies and the utilization of open-sourced data in addressing the complexities of echocardiogram interpretation, paving the way for more accurate and efficient cardiac diagnoses.

1 Introduction

Echocardiograms are one of the primary resources used by clinicians for diagnosing of cardiac conditions. They provide images and videos of a heart acquired through a low-cost, non-invasive medical procedure that uses high-frequency sound waves (ultrasound). Echocardiograms assist clinicians in the evaluation of the structure and function of the heart, including its chambers, valves, and blood flow, and at diagnosing various heart conditions, such as a heart failure, a valve disease, and congenital heart defects. During an echocardiographic study, a trained technician places on a patient's chest a device called a transducer, which emits sound waves that bounce off the heart and creates moving images. In a complete echocardiographic study, videos (referred to as views) are captured from various standard transthoracic cross sections, each of which emphasizes different areas of the heart's anatomy [41]. However, view types (also known as labels) are not assigned to these videos during the procedure. Hence, the first step in interpreting an echocardiogram is to identify the specific views that are acquired [27, 34].

Automated echocardiogram view detection is the process of using machine learning algorithms to automatically identify and specify the different views of a heart captured in echocardiograms. As part of many medical applications in



(a) Two-chamber view

(b) Four-chamber view

(c) Patches

Figure 1: Examples of echocardiographic views

cardiology, the automated identification of views is of pressing interest, as it can help improve the efficiency and accuracy of echocardiogram interpretation and medical diagnosis. Automated echocardiogram view detection poses various challenges. For instance, the acquisition of echocardiograms is a manual procedure and as such, it heavily depends on the technician's experience level. Variations within the same type of view might arise from how technicians set imaging depth, imaging gain settings, or image centering. Anatomical dissimilarities in patients or conducting the procedure with different ultrasound machines may also yield variability in the echocardiographic results [18, 48]. Moreover, while the variations within a single view can be quite large, for other views, the differences between distinct types can be rather subtle [34, 48]. Apical two-, three- and four-chamber views are a good example. These adjacent views are acquired during echocardiographic procedure by slight variations in the angle at which the technician positions the transducer [28]. Another challenge comes from the large amounts of data required to train machine learning algorithms. On one hand, there is a limited number of publicly available echocardiographic datasets. On the other hand, the use of proprietary datasets requires manual annotation which can only be provided by experts at a high cost [18, 34]. In general, the specialized nature of medical labels, as well as data privacy concerns within medical records, limit the availability of annotated medical data [3].

Several approaches have been explored to tackle automatic view detection, with convolutional neural networks (CNNs) being the go-to choice for various medical image analysis tasks. CNNs are able to automatically extract features from echocardiogram images and learn patterns in the data that allow to classify them into different views. However, all of the research on automated view detection has been conducted in a strictly supervised learning setting, with one exception exploring a semi-supervised approach. In this paper, we present the first fully unsupervised automatic echocardiographic view detection method, which avoids the need for expert manual labeling. For this, we leverage CNNs and autoencoders for self-supervised representation learning. As mentioned above, accurately discriminating between different types of views is a challenging task. In Figures 1a and 1b, examples of apical two-chamber (A2C) and apical four-chamber (A4C) views are shown. When comparing both images, it is noticeable to the human eye that certain features in them help discriminate between the two views, yet there are also elements that are common to both. In other words, there are specific details in the images that are relevant for differentiating views, while others are completely uninformative in terms of distinguishing views. To account for this, we focus on particular patches from every echocardiographic frame to use as input to our model and mask the rest of the image. Our choice of patches, shown in Figure 1c, emphasizes the regions that include discriminative information and disregards the rest.

Our unsupervised view detection framework is defined as follows. We construct embeddings from echocardiographic data by using a patch-based convolutional autoencoder. The encoder model consists of shared-weight-based CNNs for each patched input and merging by average. We further propose a trainable inverse average layer that serves as a reverse operator for the corresponding decoder. Finally, the learned representations are clustered, with each cluster corresponding to a type of an echocardiographic view. In addition, we leverage publicly available datasets to enhance the model's performance. By extending our proprietary dataset with open-source data resources, we allow the autoencoder to be trained on a wider and more diverse range of data. We conduct a thorough comparison of the outcomes of models trained on each dataset. Our experiments demonstrate a significant boost in performance when integrating public data with our proprietary data during training.

Our key contributions are as summarized as follows.

- We present the first fully unsupervised deep learning architecture for automatic echocardiogram view detection.

- We combine patches from frames and convolutional autoencoders to learn low-dimensional representations of ultrasound videos, in contrast to existing literature which does not differentiate the most discriminative information within ultrasound frames.

- We introduce a trainable inverse average layer as a method to optimize the way in which an average operation is decoded.

- We demonstrate significant improved model performance by integrating public and proprietary datasets during training.

The rest of the paper is organized as follows. In Section 2, the related work is discussed. Section 3 describes the proposed model. The computational study, data preparation and experimental results are provided in Section 4. Conclusions are given in Section 5.

2 Previous Work

2.1 Unsupervised Learning in Ultrasound Data

Representation learning for medical imaging is a well-explored technique aimed at extracting meaningful low-dimensional embeddings from data. These embeddings are subsequently employed in downstream tasks such as classification or segmentation. Specifically for ultrasound data, representation learning has been conducted using deep belief networks and proven effective at extracting features from prostate ultrasonography images to distinguish between benign and malignant lesions [2], in left ventricle segmentation [4, 5, 6], and in tongue contour extraction [23]. However, deep belief networks are complex to train, particularly when compared with more modern architectures like autoencoders, which benefit from efficient gradient-based training. Thus, autoencoders have been widely investigated as a means to obtain ultrasound embeddings by training a network to reconstruct the input data. These representations have found utility in various tasks, including segmentation of the carotid artery, midbrain, left ventricle, and prostate [35, 36], as well as diagnosing breast nodules and lesions [9]. Another application of autoencoders involves generating high-quality ultrasound images from plane wave radio frequency data [46].

More recently, contrastive representation learning has become increasingly popular. This approach entails deriving embeddings by optimizing the agreement between positive and negative pairs of data. Such representations have been assessed in supervised tasks like diagnosing lung disease [3, 8], detecting breast cancerous lesions [8], identifying prostate cancer [43], and classifying liver views [1]. It is noteworthy that contrastive learning relies on the above-mentioned pairs of positive and negative examples for training, which may not always be readily available or easy to define. In contrast, autoencoders inherently perform unsupervised learning and do not necessitate labeled data.

Previous work incorporating complementary data to support representation learning has been conducted. For instance, using narrative speech from the sonographer as an additional modality during the training process [24], accounting for anatomical information to augment the positive-negative pair sampling in contrastive learning [16] and utilizing DICOM metadata as weak labels to improve the quality of the autoencoder-generated embeddings [21]. In other cases, training models for less complex tasks and utilizing the learned weights for knowledge transferring has been explored. Examples of such tasks include the predictions of gaze [14], frame ordering, data transformations [14, 25], and data modalities [44].

As noted, it is common to use learned representations to enhance the performance of downstream supervised tasks for which labels are available. However, the fully unsupervised setting accounts for tasks in which this information is not available. For example, learning high-resolution versions of low-resolution images to improve tasks where data quality is impacting performance [32], video quality assessment modeled as an anomaly detection problem [47], and improving the estimation of displacement fields between pairs of radio-frequency data between images [10, 11].

Research on unsupervised learning focusing particularly on echocardiographic ultrasound data has also been undertaken. Automatic Otsu's thresholding has been utilized to find a closed boundary around the left atrium [38]. Contrastive-learned representations have been used to predict ejection fraction [22], left ventricular hypertrophy, and aortic stenosis [20]. Self-supervised segmentation was studied by leveraging the watershed algorithm to create weak labels [15]. Additionally, the temporal alignment of cardiac views was investigated by employing a weighted combination of temporal and spatial losses within and across views [13]. Like our approach, these prior studies emphasize unsupervised learning from echocardiographic ultrasound data. However, they do not tackle the issue of automatic unsupervised view detection.

2.2 Automatic Echocardiographic View Detection

Extensive work has been published using CNNs as backbone for supervised automatic detection of echocardiographic views [7, 17, 18, 28, 33, 48, 49]. Subsequently, knowledge distillation was used to generate lightweight models for view classification [41] and generative adversarial networks were utilized to transfer the image quality of echocardiographic images to a user-defined quality level to improve view classification [30]. Deep-learning free view classification was tacked using discriminative learning [26]. Similarly, supervised view detection has been conducted as initial step for medical diagnose pipelines for hypertrophic cardiomyopathy and cardiac amyloidosis detection [12, 45], atrial fibrillation and dilated cardiomyopathy [31], and aortic stenosis [42].

In addition to the supervised approaches, generative adversarial networks have been leveraged for view detection in a semi-supervised setting [34]. Likewise, a supervised version of a modified unsupervised deep clustering CNN [19] has been trained for view classification [40]. Although all these studies focus on echocardiographic view detection, none of them do so using a fully unsupervised approach. To the best of our knowledge, no further research has been conducted on unsupervised automatic view detection.

3 Unsupervised View Detection Framework

Our proposed unsupervised echocardiographic view detection framework consists of two parts, a patch-based convolutional autoencoder to learn frame-based embeddings and a clustering step that groups them into view-related categories.

First, the convolutional autoencoder extracts low-dimensional representations from the echocardiographic study by focusing on patches that capture the most informative sections of each frame. The obtained patch-level embeddings are further aggregated into a single frame-level representation by averaging them. However, this suggests the need of an inverse average operation that is able to decode the averaging step within the architecture. To achieve this, we introduce a trainable inverse average layer in the model. Additionally, the patch-based reconstructions are combined by a trainable reconstruction layer, designed to generate the reconstruction at a frame level. Figure 2 depicts the autoencoder's architecture. Figures 3a and 3b zoom into the construction of the customized inverse average and reconstruction layers.

Unlike many other types of real-world data, ultrasound imaging exhibits a unique characteristic in which specific regions within frames consistently contain discriminative information. These distinctive areas, often characterized by their anatomical significance or pathological features, maintain stability across consecutive frames within a given ultrasound examination. This stability facilitates the reliable identification and tracking of pertinent details. To leverage this advantage, we propose utilizing patches to learn low-dimensional representations that focus on these areas.

The selection and quantity of patches may vary, allowing for flexibility in size and shape as long as consistency is maintained across all patches per frame. Patches may or may not overlap as necessary or preferred. Overlapping can be

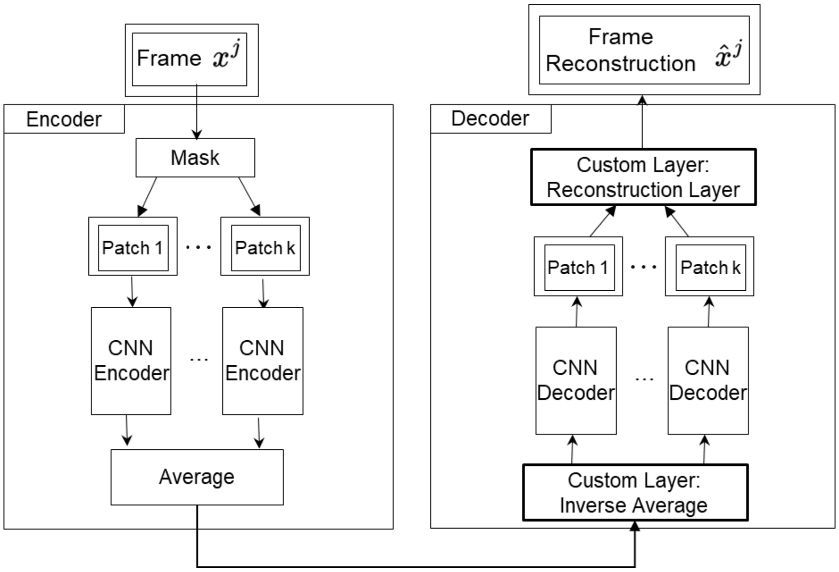
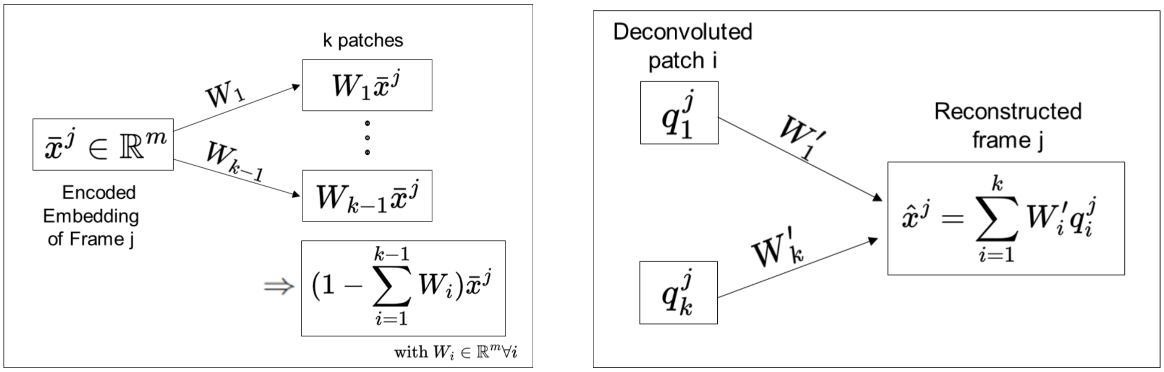


Figure 2: Patch-based convolutional autoencoder model


k patches
W_1
$W_1\bar{x}^j$
$\bar{x}^j \in \mathbb{R}^m$
W_{k-1}
$W_{k-1}\bar{x}^j$
Encoded Embedding of Frame j
$\Rightarrow$ $(1 - \sum_{i=1}^{k-1} W_i)\bar{x}^j$
with $W_i \in \mathbb{R}^m \forall i$
Deconvoluted patch i
q_1^j
W_1'
Reconstructed frame j
$\hat{x}^j = \sum_{i=1}^{k} W_i' q_i^j$
W_k'
q_k^j

(a) Inverse average layer architecture

(b) Reconstruction layer architecture

Figure 3: Architecture of trainable custom layers

advantageous to ensure that important details or features that span across adjacent patches are not missed, to provide flexibility in how specific parts of an image are represented or highlighted, and to act as a redundancy mechanism in case of frame displacement. In our study on view detection, we pinpoint six areas that highlight the discriminative features within the distinct triangular-shaped visuals of echocardiographic images. These six patches are depicted in Figure 1c.

Formally, we assume that every frame has k patches, each one of equal resolution and number of channels. In what follows, superscripts j represent a frame index and subscripts i for $i = 1, ...k$ indicate patch indexes. Let $\{c_1^j, ..., c_k^j\}$ be the set of all patches of frame x^j. A patch is represented as a tensor. Each c_i^j is encoded by a convolutional neural network CONV($c_i^j; \hat{\theta}$). Then, the embedding corresponding to frame j is defined as $\overline{x}^j = \frac{1}{k}\sum_i \text{CONV}(c_i^j; \hat{\theta})$. The decoder aims at producing a reconstructed version of x^j by reversing the operations applied to generate $\overline{x}^j$. The patch-embedding average is inversed to produce $(z_1^j, ..., z_k^j) = \text{InvAvg}(\overline{x}^j; \bar{\theta})$ as follows. Let $\bar{\theta} = [W_1 \dots W_{k-1}]$ be trainable parameters, then $z_i^j = W_i\overline{x}^j$, for $i \in \{1, 2, ...k-1\}$ and $z_k^j = (\mathbb{1} - \sum_{i=1}^{k-1} W_i)\overline{x}^j$. This clearly yields $\sum_{i=1}^{k} z_i^j = \overline{x}^j$. Then, deconvolution is used to generate patch-level representations $q_i^j = \text{DeCONV}(z_i^j; \tilde{\theta})$ for each j and i. Finally, x^j's reconstruction reads $\hat{x}^j = \sum_{i=1}^{k} \theta' q_i^j$. The network is trained by minimizing the loss function $\sum_j ||x^j - \hat{x}^j||^2$ with trainable parameters: $[\hat{\theta}, \bar{\theta}, \tilde{\theta}, \theta']$.

Upon convergence, the trained encoder is utilized to generate frame-level representations $\overline{x}^j$, which are subsequently used as inputs to train an unsupervised view detector via the K-means algorithm. The number clusters k is set to the number of echocardiographic views to detect. After clustering, frames grouped within the same cluster are anticipated to correspond to a common view. To establish video-level assignments, we identify the predominant cluster mode among all frames linked to a specific video. Subsequently, each cluster is associated with a distinct view through visual examination.

4 Computational Study

Our experimental setup consists of end-to-end training and evaluating the unsupervised view detection framework on two datasets, a relatively small proprietary dataset (US-PD) and a custom dataset built by extending the proprietary dataset with two publicly available data sources (US-PD+EXT). By utilizing this combination of proprietary and publicly available datasets, we aim to diversify the data seen by the model, thereby bolstering its generalization and robustness. Typically, obtaining unlabeled data does not pose a challenge and a wealth of data is available for model training. Hence, we optimize computational efficiency by subsampling the data.

Results from experiments conducted on both datasets are presented and compared to assess the extent to which leveraging supplementary resources enhances the downstream task of view detection. For all of the following experiments, we focus on two echocardiographic views: apical two-chamber (A2C) and apical four-chamber (A4C). Our framework operates on a fully unsupervised basis, thus presupposing the absence of labels. In order to assess our model's performance, we manually annotated a subset of 160 videos. Labels were exclusively utilized for model evaluation and not integrated into the training process (model selection is also unsupervised).

4.1 Data Preparation

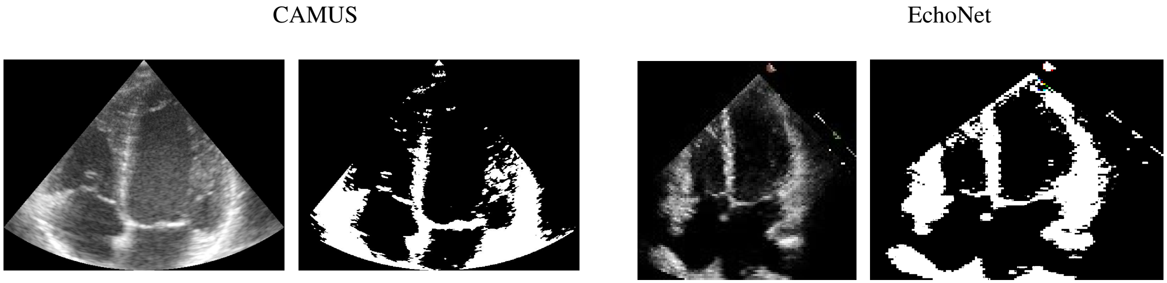
Three distinct data sources are used in the construction of datasets US-PD and US-PD+EXT. To ensure consistency, a preprocessing scheme is implemented to standardize data collected from various sources.

Propietary dataset US-PD comprises 66 videos of A2C views and 94 videos of A4C views from 100 echocardiographic patient studies. The ultrasound data was converted from the Digital Imaging and Communications in Medicine (DICOM) format to the Portable Network Graphics (PNG) format. Non-grayscale information such as the electrocardiograms and Doppler effect views were removed from the frames. Thirty percent of US-PD was allocated to the testing sample.

Extended dataset US-PD+EXT is formed by merging the following three datasets:

- US-PD,
- 250 A4C views from EchoNet [37],
- 500 A2C views and 250 A4C views from CAMUS [29].

The echocardiographic frames were resized to meet dimensions of (800,600,3) and transformed into a binary format. Figure 4 shows examples of frames before and after the preprocessing step. The resulting datasets are split into 75% and 25% for training and validation, respectively. The patient studies exhibit a wide variation in the number of frames per view (see Table 1 regarding the composition of the training data). To mitigate the data imbalance resulting from this problem, 100 frames per study were sampled for experiments using US-PD and 25 for US-PD+EXT. The final balance of each dataset is shown in Table 2. To emphasize the most informative regions within each frame, we identify six key patches and utilize them as inputs to the network. The coordinates of the six patches per frame remain fixed to those shown in Figure 1c. All of the images presented are based on the public datasets.


CAMUS

EchoNet

(a) Original frame (b) Preprocessed frame (c) Original frame (d) Preprocessed frame

Figure 4: Examples of four-chamber views from each public dataset before and after our preprocessing scheme.

View	US-PD	CAMUS	EchoNet
A2C	111	19	0
A4C	121	20	168

Table 1: Mean number of frames from each source per type of view

Dataset	A2C	A4C
US-PD (test)	38%	62%
US-PD+EXT (train)	49%	51%

Table 2: Percentage of frames per type of view

Furthermore, to visually assess the differences across the data sources, we compare their distributions by finding the closest video from each data source (US-PD, EchoNet and CAMUS) to each video in US-PD using cosine similarity as the distance metric. For simplicity, a randomly selected frame from each video is used to represent it. As noted in Figure 5, the distances between videos within US-PD are smaller than those to videos from EchoNet and CAMUS, demonstrating a clear distinction among the videos originating from separate sources.

4.2 Experimental Results

The experiments were conducted in Python using Keras 2.1.6 with Tensorflow 1.15 as the backend and ran on NVIDIA GP102 Titan Xp and Intel(R) Xeon(R) Silver 4112 CPU @ 2.60GHz. The optimization process utilized stochastic gradient descent with a learning rate of 0.001, momentum set to 0.9, norm clipped to 1, minibatch size of 1 and number of epochs of 3,000. CONV and DeCONV architectures are modified versions of VGG [39] that account for the shape of

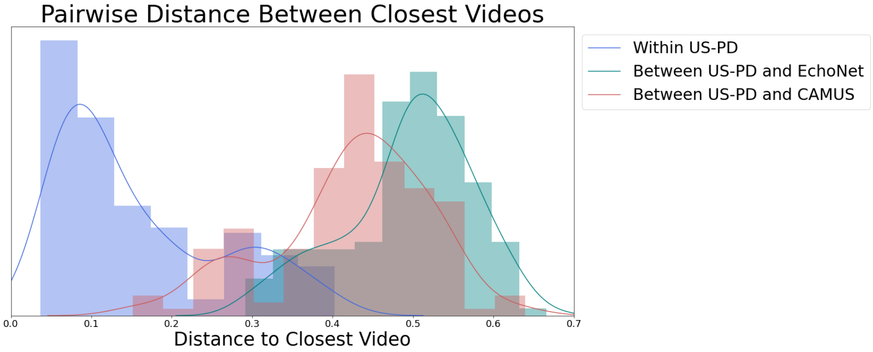


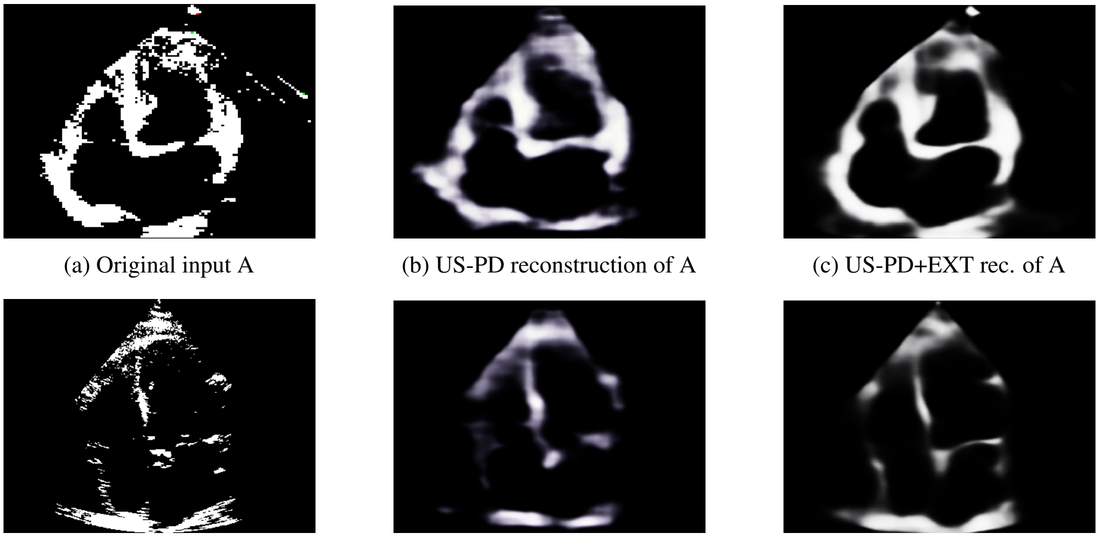
Figure 5: Distributions of the distances from each video in US-PD to its closest video in every dataset.

our input data. For simplicity, throughout the remainder of this section, we refer to the model trained using dataset US-PD as US-PD and the model trained with dataset US-PD+EXT as US-PD+EXT.

Autoencoder Models

The autoencoder models are trained to generate lower-dimensional representations of the input data and then reconstruct the original data from the compressed representation. A visual comparison of input frames from distinct A4C views, as well as their reconstructions generated by each model are shown in Figure 6.

In general, both autoencoders capture the overall pixel-based structure of the inputs. Note that the models are trained to minimize the reconstruction error, which means they aim to capture the most important features of the input data rather than producing pixel-perfect reconstructions. As a result, the reconstructed images may appear slightly blurry compared to the original input. The variations between the models can be observed in the reconstructions. For instance, in the lower left part of Figures 6d-6f, we can observe that the wall of the right atrium is better captured by US-PD+EXT than by US-PD. On the other hand, in the top row (Figures 6a-6c), the tricuspid valve is best captured by US-PD.



(a) Original input A

(b) US-PD reconstruction of A

(c) US-PD+EXT rec. of A

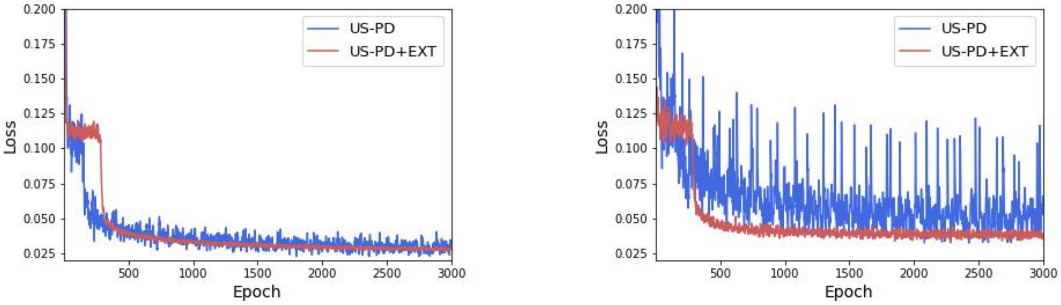
(d) Original input B

(e) US-PD reconstruction of B

(f) US-PD+EXT rec. of B

Figure 6: Comparison of two A4C view frames (A and B) and their corresponding reconstructed versions obtained by each autoencoder model. The top row depicts an image from a patient of the EchoNet dataset and the lower row a patient in the CAMUS dataset.

Figure 7 illustrates the training and validation loss curves for each model. In both Figures 7a and 7b, US-PD exhibits considerably more noise in comparison to US-PD+EXT. This suggests that the proprietary dataset may encompass noisy or inconsistent data. However, augmenting this dataset with cleaner and more stable data, even if sourced from different origins (see Figure 5), significantly contributes to stabilizing the learning process and expediting convergence.



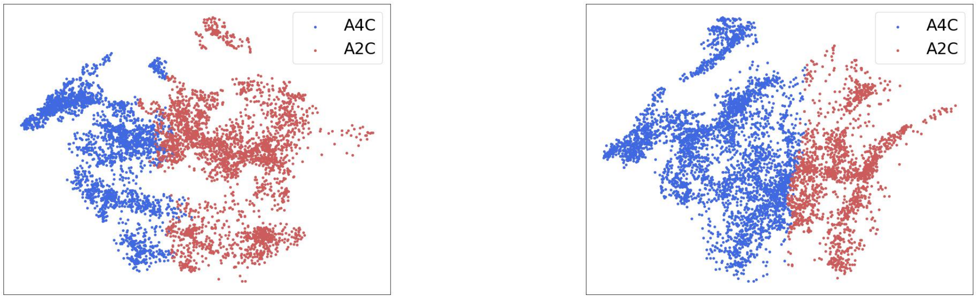
(a) Training loss

(b) Validation loss

Figure 7: Comparison of the training and validation losses for the experiments on both considered datasets

Clustering

The K-means algorithm is utilized to partition the representations into two separate clusters, where each cluster corresponds to a specific type of view. The subsequent plots are based on the test data. Figure 8 illustrates the two-dimensional projections of $\bar{x}^j$ obtained through principal component analysis (PCA). Each point denotes a frame and is color-coded according to its assigned cluster. For reference, the corresponding visualization at a video-level aggregation can be found in Appendix A) Figure 15. The metrics to evaluate the correctness of the clusters are presented in Table 3, based only on the test data. Despite being trained on a limited amount of data, US-PD achieves accurate identification of views over 60% of the time without any labels or guidance. Moreover, upon diversifying data sources, the accuracy improves by 15.55% and the F-1 score by 9%. To further assess the robustness of our approach, we derive clusters at a video level by identifying the most common cluster among the frame-level clusters of embeddings pertaining to the same ultrasound video. The corresponding results are depicted in the right-hand side of Table 3. We observe that the test metrics for US-PD are less favorable at the video-level compared to the frame-level. This poses a challenge for the view identification task, as certain frames are mistakenly assigned to the incorrect cluster. In contrast, the metrics for US-PD+EXT display consistency and slight improvements when analyzing clusters at the video-level. At this level of aggregation, the inclusion of public datasets boosts accuracy by 25.9% and F-1 by 19.8%. In summary, the performance enhancements observed in our experiments demonstrate that incorporating additional data from diverse sources into training pipelines yields large improvements.



(a) Clusters of US-PD embeddings

(b) Clusters of US-PD+EXT embeddings

Figure 8: 2D PCA projections of learned representations clustered using K-means algorithm

Ablation Study

An ablation study is conducted to understand the contributions of various components in our proposed architecture. By systematically removing or altering specific elements of the model, we observe changes in performance metrics to identify the significance of each part. We assess the impact of a) disabling the use of patches and using frames directly as input to the CNN autoencoder for representation learning, b) disabling the trainable inverse average layer and fixing

Model	Frame based		Video based	
	Accuracy	F-1	Accuracy	F-1
US-PD	0.6084	0.5978	0.5625	0.5608
US-PD+EXT	0.7030	0.6516	0.7083	0.6719
Improvement	15.55%	9.00%	25.91%	19.81%

Table 3: Test metrics per model and percentage of improvement when introducing different data sources into the training set

each patch's embedding contribution to $\frac{1}{k}$, and c) replacing the trainable weighted sum reconstruction layer with a standard sum operation. Each variation is evaluated in both datasets. The results are presented in Table 4.

The findings indicate that patches are essential for enhancing performance when diversifying data sources. Without patches, US-PD+EXT's performance is noticeably poor. Disabling the trainable inverse average and weighted sum reconstruction layers on US-PD+EXT also result in a performance degradation compared to our proposed model, though the difference is less pronounced.

Interestingly, when using a single data source (US-PD), the inclusions of patches, trainable inverse average layer, and the weighted sum reconstruction layer do not appear to significantly affect performance, as these experiments achieve results comparable to those obtained on this dataset by our proposed model (see Table 3). We attribute these results to the small size of the dataset. In summary, the ablation study validates that our proposed architecture for the US-PD+EXT model achieves superior results, underscoring the importance of diversifying data sources and leveraging the proposed architectural components to ensure high accuracy and robust performance.

	Model	Frame based $\Delta\%$		Video based $\Delta\%$	
		Accuracy	F-1	Accuracy	F-1
No Patches c_i^j	US-PD	-12.23	-9.97	0.00	0.27
	US-PD+EXT	-37.87	-33.72	-29.41	-26.76
No Inverse Average Layer $\bar{\theta}$	US-PD	-0.28	0.76	0.00	0.00
	US-PD+EXT	-0.26	-0.28	-5.87	-5.39
No Reconstruction Layer θ'	US-PD	-1.23	-0.03	0.00	-0.02
	US-PD+EXT	-2.12	-0.83	-2.94	-1.35

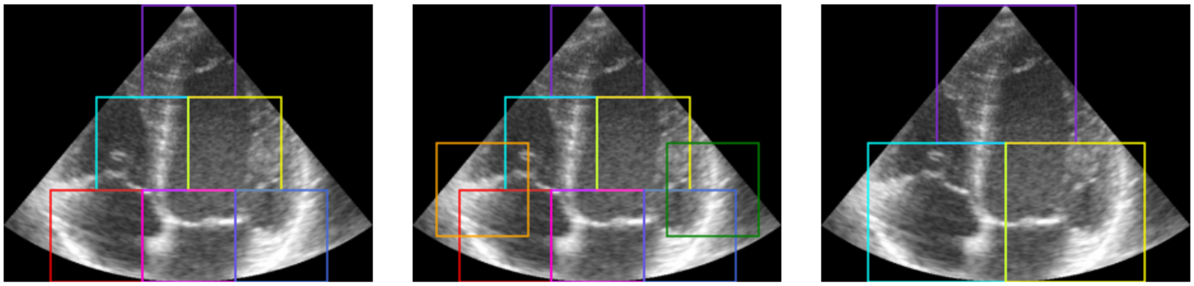
Table 4: Percentage of change in test metrics per ablated model

Patch Configuration Analysis

In this section, we explore the impact of distinct patch configurations on the performance of our model trained in both datasets. As highlighted in the ablation study, the use of patches is crucial as it allows the models to focus on the most informative parts of the ultrasound frames. We analyze the impact of using a) six non-overlapping patches (Figure 9a), b) eight patches (Figure 9b), and c) three larger patches (Figure 9c). The results, shown as the percentage of change with respect to our original patch configuration, are presented in Table 5.

Our original patch selection consistently outperforms the other tested configurations. The six non-overlapping patches result in suboptimal model performance in both datasets, likely due to the bottom patches being unable to appropriately capture the relationships within the upper and lower chambers. The configuration with eight patches performs similarly to our patch choice for US-PD but fails to generalize across diverse datasets. Among the tested configurations, the one with the three-larger-patches achieved the closest performance results to our proposed selection.

Capturing as much informative content as possible without introducing noise is crucial for enhancing model performance. Overly large patches can fail to highlight significant features distributed across the image, whereas too many patches can dilute the quality of the input data with noise. The superior performance of our patch configuration can be attributed to its ability to strike an optimal balance between information capture and noise reduction.



(a) Six non-overlapping patches

(b) Eight patches

(c) Three patches

Figure 9: Examples of different patch choices

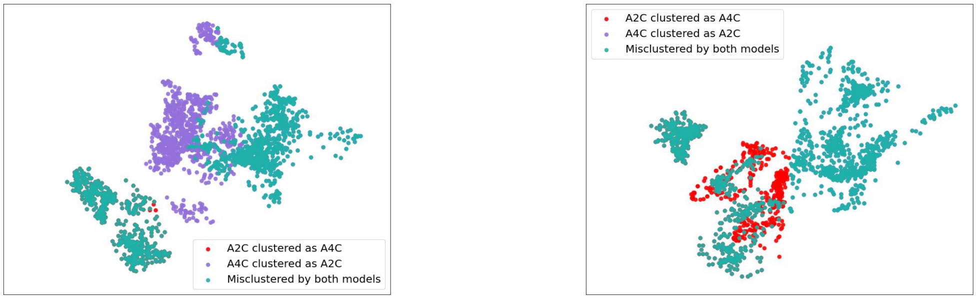
	Model	Frame based $\Delta\%$		Video based $\Delta\%$	
		Accuracy	F-1	Accuracy	F-1
Six patches with no overlap	US-PD	-29.75	-34.30	-22.22	-27.55
	UD-PD+EXT	-25.12	-19.28	-23.52	-19.50
Eight patches with overlap	US-PD	-2.30	-0.71	0.00	0.27
	UD-PD+EXT	-26.12	-20.47	-23.52	-20.46
Three patches with no overlap	US-PD	-5.74	-9.89	-11.11	-11.00
	UD-PD+EXT	-13.29	-7.01	-11.76	-6.98

Table 5: Percentage of change in test metrics per patch configuration

Misclustered Sample Analysis

An analysis comparing the misclustered embeddings was conducted to examine the differences between each model's resulting clusters. As seen in Figure 10, a significant portion of the misclustered embeddings are common to both models. This indicates that certain samples are inherently challenging to classify correctly, regardless of the model used.

However, a noteworthy observation emerges when examining the misclustered embeddings unique to each model. These clusters exhibit opposite behaviors. In Figure 10a) we observe that US-PD tends to misassign A4C frames as A2C frames. Conversely, as seen in Figure 10b), US-PD+EXT shows a tendency to miscluster A2C frames as A4C frames. These findings suggest that while both models struggle with certain frames, US-PD+EXT may have a more refined capability to accurately distinguish A4C frames, whereas US-PD performs better with A2C frames. These observations hold when analyzing the misclustered embeddings at the video level (see Figure 11).

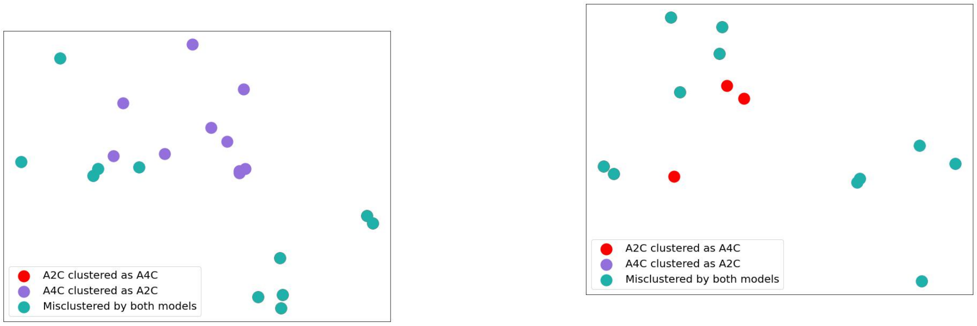


(a) Misclustered US-PD embeddings

(b) Misclustered US-PD+EXT embeddings

Figure 10: 2D PCA projections of misclustered frame embeddings

We further analyze the misclustered samples common to both models to determine the underlying reasons for their incorrect cluster assignments. We identified three main causes.



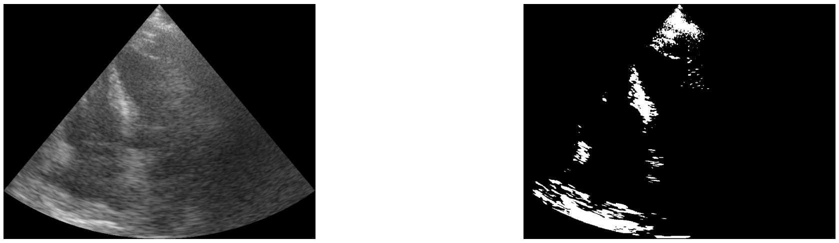
(a) Misclustered US-PD video-level embeddings

(b) Misclustered US-PD+EXT video-level embeddings

Figure 11: 2D PCA projections of misclustered embeddings aggregated per video

- Poor Quality of Sample: The raw data is characterized by low-resolution or noisy frames. These samples lack the necessary details for accurate feature extraction, leading to erroneous clustering. An example is shown in Figure 12.
- Loss of Detail During Preprocessing: The preprocessing technique, while essential for normalization and feature extraction, may inadvertently strip away essential details. Since the preprocessing employs uniform hyperparameters across all samples, there are instances where it may not perform optimally. This loss of detail poses challenges for the models to accurately distinguishing between similar frames. An example is shown in Figure 13.
- Displaced View from Standard Position: Views that are not centered or are otherwise displaced from their standard position can confuse the models. This misalignment can result in features being extracted incorrectly, leading to misclustering. An example is shown in Figure 14.

Identifying the primary factors influencing misclustered samples underscores the critical importance of utilizing reliable data to optimize model performance and provides insights to guide future improvements in data quality, preprocessing techniques, and model robustness.



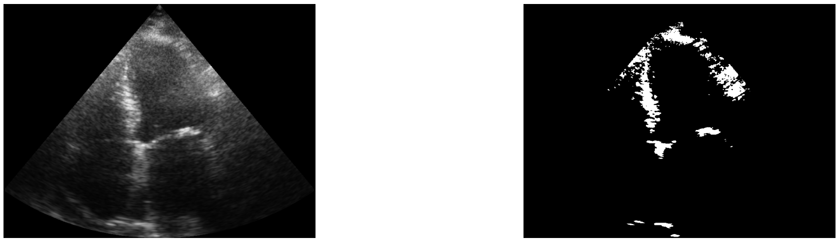
(a) Original frame

(b) Preprocessed frame

Figure 12: CAMUS A4C misclustered sample due to poor data quality

5 Conclusion

We introduce a fully unsupervised framework for automatic detection of echocardiographic views and demonstrate the feasibility of tackling this task without the need for labor-intensive labeling. Our findings highlight the advantages of integrating publicly available data sources with proprietary data both in terms of performance and training stability. Enabling our model to learn from a broader and more diverse set of data yields notable improvements in generalization. This work lays the groundwork for more efficient label-free automated pipelines for medical data and indicates strong potential for further enhancements as additional public echocardiographic datasets become accessible.



(a) Original frame (b) Preprocessed frame

Figure 13: CAMUS A4C misclustered sample due to loss of detail during preprocessing

(a) Original frame (b) Preprocessed frame

Figure 14: CAMUS A4C misclustered sample due to displacement from standard position

References

[1] A.-R. Ali, A. E. Samir, and P. Guo. Self-supervised learning for accurate liver view classification in ultrasound images with minimal labeled data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

[2] S. Azizi, F. Imani, B. Zhuang, A. Tahmasebi, J. T. Kwak, S. Xu, N. Uniyal, B. Turkbey, P. Choyke, P. Pinto, B. Wood, M. Moradi, P. Mousavi, and P. Abolmaesumi. Ultrasound-based detection of prostate cancer using automatic feature selection with deep belief networks. In *Medical Image Computing and Computer-Assisted Intervention*, 2015.

[3] S. Basu, S. Singla, M. Gupta, P. Rana, P. Gupta, and C. Arora. Unsupervised Contrastive Learning of Image Representations from Ultrasound Videos with Hard Negative Mining. In *Medical Image Computing and Computer Assisted Intervention*, 2022.

[4] G. Carneiro and J. C. Nascimento. Multiple dynamic models for tracking the left ventricle of the heart from ultrasound data using particle filters and deep learning architectures. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[5] G. Carneiro and J. C. Nascimento. Combining multiple dynamic models and deep learning architectures for tracking the left ventricle endocardium in ultrasound data. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 2013.

[6] G. Carneiro, J. C. Nascimento, and A. Freitas. The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods. *IEEE Transactions on Image Processing*, 2012.

[7] H. Chen, Q. Dou, D. Ni, J.-Z. Cheng, J. Qin, S. Li, and P.-A. Heng. Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks. In *Medical Image Computing and Computer-Assisted Intervention*, 2015.

[8] Y. Chen, C. Zhang, L. Liu, C. Feng, C. Dong, Y. Luo, and X. Wan. USCL: Pretraining deep ultrasound image diagnosis model through video contrastive representation learning. In *Medical Image Computing and Computer Assisted Intervention*, 2021.

[9] J.-Z. Cheng, D. Ni, Y.-H. Chou, J. Qin, C.-M. Tiu, Y.-C. Chang, C.-S. Huang, D. Shen, and C.-M. Chen. Computer-aided diagnosis with deep learning architecture: Applications to breast lesions in us images and pulmonary nodules in CT scans. *Scientific Reports*, 2016.

[10] R. Delaunay, Y. Hu, and T. Vercauteren. An unsupervised approach to ultrasound elastography with end-to-end strain regularisation. In *Medical Image Computing and Computer Assisted Intervention*, 2020.

[11] R. Delaunay, Y. Hu, and T. Vercauteren. An unsupervised learning approach to ultrasound strain elastography with spatio-temporal consistency. *Physics in Medicine & Biology*, 2021.

[12] R. C. Deo, J. Zhang, L. A. Hallock, S. Gajjala, L. Nelson, E. Fan, M. A. Aras, C. Jordan, K. E. Fleischmann, M. Melisko, A. Qasim, S. J. Shah, and R. Bajcsy. An end-to-end computer vision pipeline for automated cardiac function assessment by echocardiography. *ArXiv*, abs/1706.07342, 2017.

[13] F. Dezaki, C. Luong, T. Ginsberg, R. Rohling, K. Gin, P. Abolmaesumi, and T. Tsang. Echo-SyncNet: Self-supervised cardiac view synchronization in echocardiography. *IEEE Transactions on Medical Imaging*, 2021.

[14] R. Droste, Y. Cai, H. Sharma, P. Chatelain, L. Drukker, A. Papageorghiou, and J. Noble. Ultrasound image representation learning by modeling sonographer visual attention. In *Information Processing in Medical Imaging*, 2019.

[15] D. L. Ferreira, Z. Salaymang, and R. Arnaout. Self-supervised learning for label-free segmentation in cardiac ultrasound. *ArXiv*, abs/2210.04979, 2022.

[16] Z. Fu, J. Jiao, R. Yasrab, L. Drukker, A. T. Papageorghiou, and J. A. Noble. Anatomy-aware contrastive representation learning for fetal ultrasound. *European Conference on Computer Vision*, 2022.

[17] X. Gao, W. Li, M. Loomes, and L. Wang. A fused deep learning architecture for viewpoint classification of echocardiography. *Information Fusion*, 2017.

[18] Y. Gao, Y. Zhu, B. Liu, Y. Hu, G. Yu, and Y. Guo. Automated recognition of ultrasound cardiac views based on deep learning with graph constraint. *Diagnostics*, 2021.

[19] X. Guo, X. Liu, E. Zhu, and J. Yin. Deep clustering with convolutional autoencoders. In *Neural Information Processing*, 2017.

[20] G. Holste, E. K. Oikonomou, B. J. Mortazavi, Z. Wang, and R. Khera. Self-supervised contrastive learning of echocardiogram videos enables label-efficient cardiac disease diagnosis. *ArXiv*, abs/2207.11581, 2023.

[21] S.-Y. Hu, S. Wang, W.-H. Weng, J. Wang, X. Wang, A. Ozturk, Q. Li, V. Kumar, and A. E. Samir. Self-supervised pretraining with DICOM metadata in ultrasound imaging. In *Machine Learning for Healthcare*, 2020.

[22] Y. Hu, T. M. Sutter, E. Ozkan, and J. E. Vogt. Self-supervised learning to predict ejection fraction using motion-mode images. In *International Conference on Learning Representations, Workshop on Machine Learning & Global Health*, 2023.

[23] A. Jaumard-Hakoun, K. Xu, P. Roussel-Ragot, G. Dreyfus, and B. Denby. Tongue contour extraction from ultrasound images based on deep neural network. *ArXiv*, abs/1605.05912, 2016.

[24] J. Jiao, Y. Cai, M. Alsharid, L. Drukker, A. T. Papageorghiou, and J. A. Noble. Self-supervised contrastive video-speech representation learning for ultrasound. In *Medical Image Computing and Computer-Assisted Intervention*, 2020.

[25] J. Jiao, R. Droste, L. Drukker, A. Papageorghiou, and J. Noble. Self-supervised representation learning for ultrasound video. In *IEEE International Symposium on Biomedical Imaging*, 2020.

[26] H. Khamis, G. Zurakhov, V. Azar, A. Raz, Z. Friedman, and D. Adam. Automatic apical view classification of echocardiograms using a discriminative learning dictionary. *Medical Image Analysis*, 2017.

[27] K. Kusunose. Steps to use artificial intelligence in echocardiography. *Journal of Echocardiography*, 2020.

[28] K. Kusunose, A. Haga, M. Inoue, D. Fukuda, H. Yamada, and M. Sata. Clinically feasible and accurate view classification of echocardiographic images using deep learning. *Biomolecules*, 2020.

[29] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, C. Lartizien, J. D'hooge, L. Løvstakken, and O. Bernard. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. *IEEE Transactions on Medical Imaging*, 2019.

[30] Z. Liao, M. H. Jafari, H. Girgis, K. Gin, R. Rohling, P. Abolmaesumi, and T. Tsang. Echocardiography view classification using quality transfer star generative adversarial networks. In *Medical Image Computing and Computer Assisted Intervention*, 2019.

[31] X. Liu, Y. Fan, S. Li, M. Chen, M. Li, W. K. Hau, H. Zhang, L. Xu, and A. P.-W. Lee. Deep learning-based automated left ventricular ejection fraction assessment using 2-D echocardiography. *American Journal of Physiology-Heart and Circulatory Physiology*, 2021.

[32] J. Lu and W. Liu. Unsupervised super-resolution framework for medical ultrasound images using dilated convolutional neural networks. In *IEEE International Conference on Image, Vision and Computing*, 2018.

[33] A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout. Fast and accurate view classification of echocardiograms using deep learning. *Digital Medicine*, 2018.

[34] A. Madani, J. Ong, A. Tibrewal, and M. Mofrad. Deep echocardiography: Data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. *npj Digital Medicine*, 2018.

[35] R.-M. Menchón-Lara and J. L. Sancho-Gómez. Fully automatic segmentation of ultrasound common carotid artery images based on machine learning. *Neurocomputing*, 2015.

[36] F. Milletari, S.-A. Ahmadi, C. Kroll, C. Hennersperger, F. Tombari, A. Shah, A. Plate, K. Boetzel, and N. Navab. Robust segmentation of various anatomies in 3d ultrasound using hough forests and learned data representations. In *Medical Image Computing and Computer-Assisted Intervention*, 2015.

[37] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. Liang, E. A. Ashley, and J. Y. Zou. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, 2020.

[38] K. T. Shahid and I. Schizas. Unsupervised mitral valve tracking for disease detection in echocardiogram videos. *Journal of Imaging*, 2020.

[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[40] J. Tromp, P. Seekings, C.-L. Hung, M. Iversen, M. Frost, W. Ouwerkerk, Z. Jiang, F. Eisenhaber, R. Goh, H. Zhao, W. Huang, L.-H. Ling, D. Sim, P. Cozzone, A. Richards, H. Lee, S. Solomon, C. Lam, and J. Ezekowitz. Automated interpretation of systolic and diastolic function on the echocardiogram: a multicohort study. *The Lancet Digital Health*, 2022.

[41] H. Vaseli, Z. Liao, A. H. Abdi, H. Girgis, D. Behnami, C. Luong, F. T. Dezaki, N. Dhungel, R. Rohling, K. Gin, P. Abolmaesumi, and T. Tsang. Designing lightweight deep learning models for echocardiography view classification. In *Medical Imaging: Image-Guided Procedures, Robotic Interventions, and Modeling*, 2019.

[42] B. S. Wessler, Z. Huang, G. M. Long, S. Pacifici, N. Prashar, S. Karmiy, R. A. Sandler, J. Z. Sokol, D. B. Sokol, M. M. Dehn, L. Maslon, E. Mai, A. R. Patel, and M. C. Hughes. Automated detection of aortic stenosis using machine learning. *Journal of the American Society of Echocardiography*, 2023.

[43] P. F. R. Wilson, M. Gilany, A. Jamzad, F. Fooladgar, M. N. N. To, B. Wodlinger, P. Abolmaesumi, and P. Mousavi. Self-supervised learning with limited labeled data for prostate cancer detection in high-frequency ultrasound. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 2023.

[44] Z. Xiang, Q. Zhuo, C. Zhao, X. Deng, T. Zhu, T. Wang, W. Jiang, and B. Lei. Self-supervised multi-modal fusion network for multi-modal thyroid ultrasound image diagnosis. *Computers in Biology and Medicine*, 2022.

[45] J. Zhang, S. Gajjala, P. Agrawal, G. Tison, L. Hallock, L. Beussink, M. Lassen, E. Fan, M. Aras, C. Jordan, K. Fleischmann, M. Melisko, A. Qasim, S. Shah, R. Bajcsy, and R. Deo. Fully automated echocardiogram interpretation in clinical practice: Feasibility and diagnostic accuracy. *Circulation*, 2018.

[46] J. Zhang, Q. He, Y. Xiao, H. Zheng, C. Wang, and J. Luo. Ultrasound image reconstruction from plane wave radio-frequency data by self-supervised deep neural network. *Medical Image Analysis*, 2021.

[47] H. Zhao, Q. Zheng, C. Teng, R. Yasrab, L. Drukker, A. T. Papageorghiou, and J. A. Noble. Towards unsupervised ultrasound video clinical quality assessment with multi-modality data. In *Medical Image Computing and Computer Assisted Intervention*, 2022.

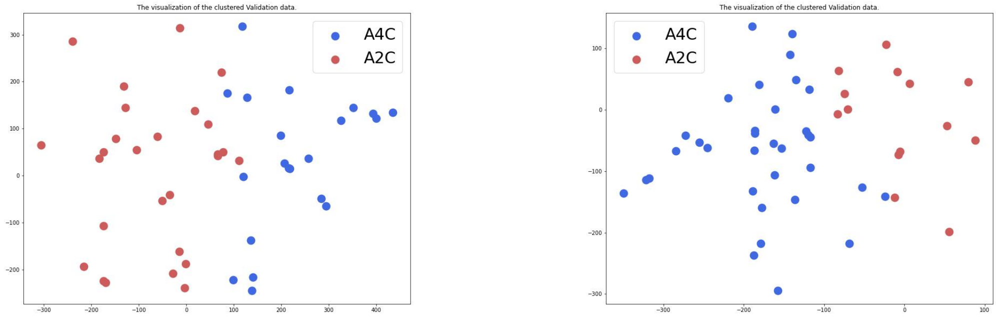
[48] Y. Zhu, J. Ma, Z. Zhang, Y. Zhang, S. Zhu, M. Liu, Z. Zhang, C. Wu, X. Yang, J. Cheng, D. Ni, M. Xie, W. Xue, and L. Zhang. Automatic view classification of contrast and non-contrast echocardiography. *Frontiers in Cardiovascular Medicine*, 2022.

[49] A. Østvik, E. Smistad, S. A. Aase, B. O. Haugen, and L. Lovstakken. Real-time standard view classification in transthoracic echocardiography using convolutional neural networks. *Ultrasound in Medicine & Biology*, 2019.

Appendix A) Clusters per echocardiographic video

Figure 15 illustrates the principal component analysis (PCA) two-dimensional projections of the video-level aggregations of $\bar{x}^j, \forall j$ corresponding to the same video, which are obtained by averaging. Each point denotes a video and is color-coded according to its assigned cluster.

The clusters displayed in this figure yield the video-based test metrics for each model that are shown in Table 3. Figure 8 illustrates the equivalent visualization of the cluster assignments at a frame level.



(a) Clusters of US-PD embeddings

(b) Clusters of US-PD+EXT embeddings

Figure 15: 2D PCA projections of video-level aggregated embeddings and its corresponding clusters