Interpretable Anomaly Detection in Echocardiograms with Dynamic Variational Trajectory Models

Alain Ryser [1] **Laura Manduchi** [1] **Fabian Laumer** [1] **Holger Michel** [2] **Sven Wellmann** [2] **Julia E. Vogt** [1]

Abstract

We propose a novel anomaly detection method for echocardiogram videos. The introduced method takes advantage of the periodic nature of the heart cycle to learn different variants of a *variational latent trajectory* model (TVAE). The models are trained on the healthy samples of an in-house dataset of infant echocardiogram videos consisting of multiple chamber views to learn a normative prior of the healthy population. During inference, maximum a posteriori (MAP) based anomaly detection is performed to detect out-of-distribution samples in our dataset. The proposed method reliably identifies severe congenital heart defects, such as Ebstein's Anomaly or Shone-complex. Moreover, it achieves superior performance over MAP-based anomaly detection with standard variational autoencoders on the task of detecting pulmonary hypertension and right ventricular dilation. Finally, we demonstrate that the proposed method provides interpretable explanations of its output through heatmaps which highlight the regions corresponding to anomalous heart structures.

1. Introduction

Congenital heart defects (CHDs) account for about 28% of all congenital defects worldwide (Van Der Linde et al., 2011). CHDs manifest in several different heart diseases with various degrees of frequency and severity, and are usually diagnosed primarily with echocardiography. Echocardiography is one of the most common non-invasive screening tools due to the rapid data acquisition, low cost, portability, and measurement without ionizing radiation. Early screening of heart defects in newborns is crucial to ensure the long-term health of the patient (Buskens et al., 1996; Singh & McGeoch, 2016; Van Velzen et al., 2016). However, due to the subtlety of various heart defects and the inherently noisy nature of echocardiogram video (echo) data, a thorough examination of the heart and the diagnosis of CHD remains a challenging and time-consuming process, raising the need for an automated approach. Still, collecting real-world datasets from large populations to apply state-of-the-art supervised deep learning methods is often infeasible. The reason is that many CHDs like Ebstein's Anomaly, Shone-complex, or complete atrioventricular septal defect (cAVSD) rarely occur, making the dataset extremely imbalanced. On the other hand, we have access to an abundance of echos from healthy infant hearts generated during standard screening procedures, often performed on infants shortly after birth.

In this work, we introduce a novel anomaly detection method to identify a variety of CHDs. The proposed approach learns a structured normative prior of healthy newborn echos using a periodic variational latent trajectory model. At test time, the method can detect out-of-distribution samples corresponding to CHDs. The advantage of this approach is that the model is trained purely on healthy samples, eliminating the need to collect large amounts of often rarely occurring CHDs.

In anomaly detection, we assume that all data is drawn from a space $\mathcal{X}$ with some probability density p_H. Anomalies are then defined to be samples drawn from low probability regions of $\mathcal{X}$ under p_H. More formally, an anomaly space $\mathcal{A} \subset \mathcal{X}$ under density p_H and anomaly threshold $\tau \geq 0$ is defined by

$$\mathcal{A} = \{x \in \mathcal{X}; p_H(x) \leq \tau\}$$

Note that τ is a task-specific measure, as the definition of anomaly can vary drastically over different problem settings. Consequently, most anomaly detection algorithms assign anomaly scores rather than discriminating between normal and anomalous samples.

In this work we focus on reconstruction-based approaches, which encompass some of the most widespread methods for anomaly detection (Chalapathy & Chawla, 2019; Ruff et al., 2021; Pang et al., 2021). This family of methods aims to

[1]Department of Computer Science, ETH Zürich [2]Department of Neonatology, University Children's Hospital Regensburg (KUNO), University of Regensburg, Germany. Correspondence to: Alain Ryser <alain.ryser@inf.ethz.ch>.

Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML).

learn generative models that can reconstruct normal samples well but decrease in performance for anomalous inputs. A given measure $\alpha_f(x)$ that quantifies the reconstruction quality achieved by model f when given sample x can then be interpreted as the anomaly score of x. The models are commonly trained on healthy samples, and during inference, an anomalous sample x_a is assumed to get projected into the learned normal latent space. This effectively leads to high reconstruction errors, resulting in high anomaly scores $\alpha_f(x_a)$. More recently, (Chen et al., 2020) proposed a variation of the reconstruction-based approach that allows us to incorporate prior knowledge on anomalies during inference by detecting anomalies using a *maximum a posteriori* (MAP) based approach. However, this approach requires an estimate of the log-likelihood, which restricts model choice to generative models such as *variational autoencoders* (VAE Kingma & Welling (2013)).

Although various generative architectures have been proposed in the literature, little effort has been directed toward echocardiogram videos. One exception is the work of Laumer et al. (2020), where the authors introduced a model that specifically targets the periodicity of heartbeats for ejection fraction prediction and arrhythmia classification. However, the model enforces rather restrictive assumptions on the heart dynamics and is purely deterministic in nature. In contrast, we propose a variational latent trajectory model that overcomes the simplistic assumptions of previous approaches and learns a distribution over dynamic trajectories, enabling the detection of different types of CHDs in echocardiograms using the MAP approach. Furthermore, the proposed algorithm allows us to explain predictions by producing heatmaps that highlight regions corresponding to detected anomalies, which ultimately helps clinicians in building trust in the proposed approach. We provide the code to our models on GitHub[1].

To summarize, the contributions of this paper are the following:

1. We propose a novel variational latent trajectory model (TVAE) for reconstruction-based anomaly detection on echocardiogram videos.

2. We perform extensive evaluation of the proposed method on the challenging task of CHD detection in a real-world dataset.

3. We complement our predictions with decision heatmaps, which highlight the regions of the echocardiograms corresponding to anomalous heart structures.

[1] https://github.com/alain-ryser/tvae

2. Related Work

The rapid data acquisition, the high observer variation in their interpretation, and the non-invasive technology have made echocardiography a suitable data modality for an abundance of machine learning algorithms. In recent years, a variety of algorithms for *segmentation* (Dong et al., 2016; Moradi et al., 2019; Leclerc et al., 2019), *view classification* (Gao et al., 2017; Vaseli et al., 2019) or *disease prediction* (Madani et al., 2018; Kwon et al., 2019) have been proposed. However, their performance often relies on the assumption that a large *labeled* dataset can be collected. This assumption does not hold for rare diseases, where the amount of collected data is often too scarce to train a supervised algorithm. Hence, reconstruction-based anomaly detection algorithms could be used in such a setting, but their application to echocardiography is, to the best of our knowledge, left unexplored.

Previous work on reconstruction based anomaly detection are often based on generative models, such as *autoencoders* (AE) (Principi et al., 2017; Chen et al., 2017; Chen & Konukoglu, 2018; Pawlowski et al., 2018) or *variational autoencoders* (VAE Kingma & Welling (2013)) (An & Cho, 2015; Park et al., 2018; Xu et al., 2018; Cerri et al., 2019; You et al., 2019). Their application to the medical domain is mostly limited to disease detection in MRI (Baur et al., 2018; Chen & Konukoglu, 2018; Baur et al., 2020; Chen et al., 2020; Baur et al., 2021; Pinaya et al., 2021), where anomalies are often easily detectable as they are clearly defined by regions of tissue that contain lesions. On the other hand, pathologies of CHDs in echos are largely heterogeneous and usually cannot be described by unique structural differences from healthy echos. Identifying them is often challenging, as they can be caused by small perturbations of ventricles (ventricular dilation) or subtle malfunctions like pressure differences between chambers in certain phases of the cardiac cycle (pulmonary hypertension). Detecting certain CHDs thus requires the inclusion of temporal structures in addition to the spatial information leveraged in MRI anomaly detection.

Different extensions to AE/VAE have been proposed to perform reconstruction-based anomaly detection methods on video data (Xu et al., 2015; Hasan et al., 2016; Yan et al., 2018). However, these methods are often designed for abnormal event detection, where anomalies can arise and disappear throughout the video. On the other hand, we are interested in whether a given video represents a healthy or anomalous heart. Another method for video anomaly detection is *future frame prediction* (Liu et al., 2018), which trains models to predict a video frame from one or more previous ones. During inference, it is assumed that such a model achieves better performance on normal than on anomalous frames. Recently, (Yu et al., 2020) proposed

a method that combines reconstruction and future frame prediction-based approaches in one framework. Though achieving good performance on videos with varying scenes, future frame prediction does not seem suitable for echos as returning any input frame will always lead to good prediction scores due to the periodic nature of the cardiac cycle. An entirely different approach to anomaly detection is given by *One-Class Classification* (Moya & Hush, 1996). In contrast to the previous approaches, the latter relies on discriminating anomalies from normal samples instead of assigning an anomaly score. This is usually achieved by learning a high-dimensional manifold that encloses normal data. The surface of this manifold then serves as a decision boundary that discriminates anomalies from normal samples. One of the more prominent methods of that family is the so-called *Support Vector Data Description* (SVDD) (Tax & Duin, 2004) model. The SVDD learns parameters of a hypersphere that encloses the training data. Similar to SVMs, it provides a way to introduce some slack into the estimation process, allowing certain normal samples to lie outside the decision boundary. A similar approach is given by the *One-Class SVMs* (OC-SVM) (Schölkopf et al., 2001), where anomalies are discriminated from normal samples by learning a hyperplane instead of a hypersphere. Like with SVMs, the expressivity of SVDD and OC-SVM can be drastically improved by introducing kernelized versions (Ratsch et al., 2002; Ghasemi et al., 2012; Dufrenois, 2014; Gautam et al., 2019). More recently, deep neural networks have been proposed to perform anomaly detection based on similar principles (Ruff et al., 2018; Sabokrou et al., 2018; Ruff et al., 2020; Ghafoori & Leckie, 2020). While conceptually interesting, One-Class Classification methods often require large amounts of data to work accurately, making them unsuitable in many clinical applications.

3. Methods

In this work, we propose a probabilistic latent trajectory model to perform reconstruction-based anomaly detection on echocardiogram videos. We take inspiration from latent trajectory models (Louis et al., 2019; Laumer et al., 2020) and introduce the trajectory variational autoencoder (TVAE), which learns a structured normative distribution of the heart's shape and dynamic. In particular, the model encodes the echos into stochastic trajectories in the latent space of a VAE, enabling us to accurately generate high-quality reconstructions while maintaining a low dimensional latent bottleneck. We present three different TVAE variants. The TVAE-C and TVAE-R leverage trajectories that assume strict periodic movements of the heart, while TVAE-S is more general and allows shifts in the spatial representation throughout the video, improving the quality of the normative prior. The learned approximate distribution of healthy hearts then allows us to detect anomalies post-hoc using a maximum a posteriori (MAP) approach (Chen et al., 2020). High-quality normative reconstructions and informative latent representations are essential to detect out-of-distribution echos correctly.

3.1. Latent Trajectory Model

The latent trajectory model (Laumer et al., 2020) is an autoencoder that is designed to learn latent representations from periodic sequences of the heart, i.e. echos in this case. The main idea is to capture the periodic nature of the observed data by learning an encoder ϕ that maps an echo $X := (\vec{x}^{(j)}, t^{(j)})_{j=1}^{T}$ with frames $\vec{x}^{(j)} \in \mathbb{R}^{w \times h}$ at time points $t^{(j)}$ to a prototypical function $\vec{\ell}_{circular}(t; \phi(X))$ whose parameters contain information about the heart's shape and dynamic. The decoder ψ reconstructs the original video frame by frame from the latent embedding $\vec{\ell}_{circular}$ with $\tilde{\vec{x}}^{(j)} = \psi(\vec{\ell}_{circular}(t^{(j)}; \phi(X)))$. Here, $\vec{\ell}_{circular}$ corresponds to the following cyclic trajectory:

$$\vec{\ell}_{circular}(t; f, \omega, \vec{b}) = \begin{pmatrix} \cos(2\pi f t - \omega) + b_1 \\ \sin(2\pi f t - \omega) + b_2 \\ b_3 \\ \vdots \\ b_d \end{pmatrix},$$

where the frequency parameter, $f > 0$, corresponds to the number of cycles per time unit, and the offset parameter $\omega \in [0, 2\pi]$ allows the sequence to start at an arbitrary point within the (cardiac) cycle. The parameter $\vec{b} \in \mathbb{R}^d$ characterizes the *spatial information* of the signal. This model thus describes a simple tool to learn the disentanglement of temporal components (f, ω) from a common spatial representation $(\vec{b})$ for a given echo. On the other hand, the assumptions made may be too simplistic to result in good reconstructions. We will address this issue in the following sections.

3.2. Dynamic Trajectories

The above formulation, $\vec{\ell}_{circular}$, allows modeling time-related information only through the first two latent dimensions, thereby limiting the amount of time-dependent information that can be encoded in the latent space. The reduced flexibility results in insufficient reconstruction quality, impairing the reconstruction-based anomaly detection performance. To circumvent this problem, we distribute time-dependent components over each dimension of the latent space while retaining the periodicity. We thus define the rotated trajectory function $\vec{\ell}_{rot}$ as

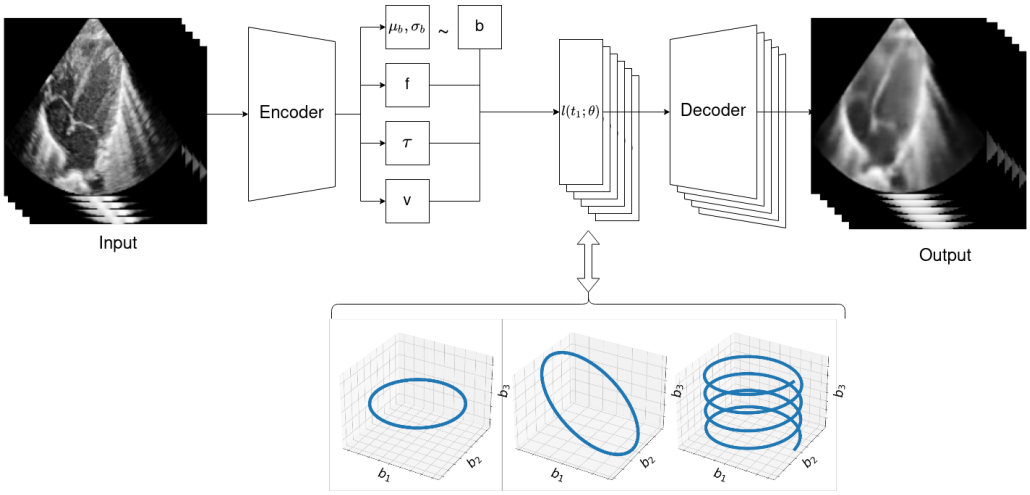


Figure 1. Overview of the model architecture with $\vec{\ell}_{circular}$ (left), $\vec{\ell}_{rot}$ (middle) and $\vec{\ell}_{spiral}$ (right).

$$\vec{\ell}_{rot}(t; f,\omega, \vec{b}) = \begin{pmatrix} \cos(2\pi ft - \omega) - \sin(2\pi ft - \omega) + b^{(1)} \\ \cos(2\pi ft - \omega) + \sin(2\pi ft - \omega) + b^{(2)} \\ \vdots \\ \cos(2\pi ft - \omega) + \sin(2\pi ft - \omega) + b^{(d)} \end{pmatrix}.$$

Furthermore, in real-world applications, it is often the case that doctors change certain settings of the echocardiogram machine during screening to get better views of certain cardiac structures. Additionally, some patients might slightly move while scans are performed, which leads to a displacement of the heart with respect to the transducer position throughout an echo recording. This is particularly prominent in our in-house dataset, which consists of echocardiograms of newborn children. Such echocardiograms are not necessarily well represented with a simple periodic trajectory. Over multiple cycles, the spatial structure of a sample shifts and looks different than in the beginning, even though temporal information like the frequency or phase shift is preserved. The trajectory model described by $\vec{\ell}_{rot}$ thus fails in such scenarios, which can manifest in two different ways: either the model incurs a local optima with high reconstruction error, or the model reconstructs the video from one long cycle, hence not leveraging the heart cycle periodicity. Thus, to account for movements of the recording device, we extend $\vec{\ell}_{rot}$ with a velocity parameter $v \in \mathbb{R}$ that allows the model to learn gradual shifts of the latent trajectory over time, resulting in a trajectory that is no longer circular but a spiral embedded in high dimensional space. More formally, we define the spiral trajectory function as

$$\vec{\ell}_{spiral}(t; f, \omega, v, \vec{b})_i = \vec{\ell}_{rot}(t; f, \omega, \vec{b})_i + tv$$

3.2.1. Variational Formulation

Previous work often applied VAEs to anomaly detection, as its generative nature enables more sophisticated variants of reconstruction-based anomaly detection (Baur et al., 2018; Xu et al., 2018; Chen et al., 2020). Thus, we extend the presented model with a stochastic layer and introduce the variational latent trajectory model.

We modify the encoder $\phi(X;\theta)$ such that it outputs trajectory parameters $v, f, \omega \in \mathbb{R}$ and $\vec{\mu_b}, \vec{\sigma_b} \in \mathbb{R}^d$. The model is then extended with a stochastic layer by defining $\vec{b} \sim q_\theta(\cdot|X) := \mathcal{N}(\vec{\mu_b}, diag(\vec{\sigma_b}))$. While we aim to learn a distribution over heart shapes, we would also like to accurately identify the frequency f, phase shift ω, and spatial shift v given an echo video X, instead of sampling them from a latent distribution. We thus leave those parameters deterministic. Next, we define an isotropic Gaussian prior $p(\vec{b}) := \mathcal{N}(0, \mathbb{I})$ on $\vec{b}$ and assume that

$$\begin{aligned} x^{(i)} &\sim p_\eta(X|\vec{b}, f, \omega, v) \\ &:= \mathcal{N}(\psi(\vec{\ell}_{spiral}(t^{(i)}; f, \omega, v, \vec{b}); \eta), \sigma\mathbb{I}), \end{aligned}$$

where ψ is our decoder with weights η and σ is some fixed constant. Given these assumptions, we are able to derive the following *evidence lower bound* (ELBO):

$$\begin{aligned} ELBO(X) &:= \\ &E_{q_\theta(\vec{b}|X)}[\log(p_\eta(X|\vec{b}, \phi_f(X), \phi_\omega(X), \phi_v(X)))] \\ &- KL[q_\theta(\vec{b}|X)||p(\vec{b})]. \end{aligned}$$

Here, $\phi_f(X)$, $\phi_\omega(X)$ and $\phi_v(X)$ are the trajectory parameter outputs of the encoder ϕ for f, ω and v, respectively. Note that VAEs on $\vec{\ell}_{circular}$ and $\vec{\ell}_{rot}$ are defined in a similar fashion. A derivation of this ELBO can be found in Appendix A.

3.2.2. ANOMALY DETECTION

The variational formulation of the latent trajectory model allows us to perform anomaly detection by *Maximum a Posteriori* (MAP) inference as proposed in Chen et al. (2020). The authors suggest that anomalies can be modeled as an additive perturbation of a healthy sample. Following their reasoning we define:

$$X_H \coloneqq (\vec{x}_H^{(j)}, t^{(j)})_{j=1}^T \sim \mathcal{H}$$
$$Y \coloneqq (\vec{y}^{(j)}, t^{(j)})_{j=1}^T \sim \mathcal{D}$$
$$A \coloneqq (\vec{a}^{(j)}, t^{(j)})_{j=1}^T,$$

where $\mathcal{H}$ is the healthy data distribution, $\mathcal{D}$ the overall data distribution (i.e., including anomalies) and A the anomalous perturbation. We then assume that

$$\vec{y}^{(j)} = \vec{x}_H^{(j)} + \vec{a}^{(j)}.$$

In the case of CHD, A could, e.g., remove walls between heart chambers or produce holes in the myocardium for specific frames. The anomaly score α can then be defined as $\alpha(Y) \coloneqq \frac{1}{T}\sum_{j=1}^T \|\vec{a}^{(j)}\|_2^2$. When training on healthy samples only, i.e. $\vec{a}^{(j)} = 0$ for all $j \in \{1, ..., T\}$, the variational latent trajectory model learns to approximate $P(X_H)$ by maximizing $ELBO(X_H)$. We then use the MAP estimation to approximate the posterior distribution of X_H given a sample Y. Hence, by using Bayes' theorem

$$P(X_H|Y) \propto P(Y|X_H)P(X_H),$$

we can estimate X_H as follows:

$$\tilde{X}_H = \arg\max_{X_H}(\log(P(Y|X_H)) + ELBO(X_H)),$$

where we use the concavity of the logarithm and the fact that $\log(P(X_H)) \geq ELBO(X_H)$. Next, we compute $\vec{\tilde{a}}^{(j)} = \vec{y}^{(j)} - \vec{\tilde{x}}_H^{(j)}$ and calculate the anomaly score as $\alpha(Y) \coloneqq \frac{1}{T}\sum_{t=1}^T \|\vec{\tilde{a}}^{(t)}\|_2^2$. Similar to (Chen et al., 2020), we choose $\log P(Y|X) = \|(\vec{x}^{(j)} - \vec{y}^{(j)})_{j=1}^T\|_{TV}$, where $\|\cdot\|_{TV}$ denotes the *Total Variation norm* in ℓ_1, as this leverages the assumption that anomalies should consist of contiguous regions rather than single pixel perturbations. Note that since we have a temporal model, we can incorporate temporal gradients into the TV norm, i.e.,

$$\|X\|_{TV} \coloneqq \sum_{i=1}^w \sum_{j=1}^h \sum_{k=1}^T \|\nabla \vec{x}_{ij}^{(k)}\|_1.$$

In our experiments, we approximate gradients by

$$\nabla \vec{x}_{ij}^{(k)} \approx \begin{pmatrix} \vec{x}_{(i+1)j}^{(k)} - \vec{x}_{(i-1)j}^{(k)} \\ \vec{x}_{i(j+1)}^{(k)} - \vec{x}_{i(j-1)}^{(k)} \\ \vec{x}_{ij}^{(k+1)} - \vec{x}_{ij}^{(k-1)} \end{pmatrix}.$$

4. Experiments

All experiments are conducted on a novel in-house dataset of echocardiograms of newborns. We perform three separate anomaly detection tasks, namely detecting severe structural defects (SSD), right ventricular dilation (RVDil), or pulmonary hypertension (PH). For each task, we define samples that do not contain the respective lesion as part of the normal distribution. We perform all anomaly detection tasks on both the apical four-chamber (4CV) and the parasternal long-axis (PLAX) view. Videos were preprocessed and resampled such that they consist of 25 frames. More details on the collected dataset and preprocessing can be found in Appendix B.

In addition to the variational latent trajectory models with the circular (TVAE-C), rotated (TVAE-R), and spiral (TVAE-S) trajectories described in Section 3.1, we train a standard variational autoencoder (Kingma & Welling, 2013) model on the individual video frames of the dataset as a baseline. We present an outline of the model architecture in Figure 1 and refer to Appendix C for a more detailed description of the network.

We run experiments by training the models exclusively on samples that do not contain the corresponding CHD to learn the normative prior. Each experiment is trained on 10 separate data splits, where we leave out the respective anomalous samples, 30 healthy samples for PH and RVDil, and 7 healthy samples for SSD to evaluate on test time.

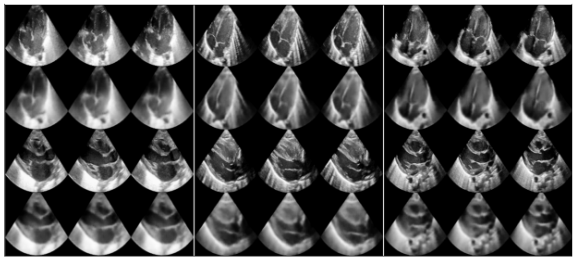
4.1. Reconstruction

The reconstruction quality is directly related to reconstruction-based anomaly detection performance, as we rely on the *manifold* and *prototype* assumptions formalized in (Ruff et al., 2021). The manifold assumption is often used in many machine learning-based applications and states that $\mathcal{X}$, the space of healthy echos, can be generated from some latent space $\mathcal{Z}$ by a decoding function ψ and that it is possible to learn a function ϕ that encodes $\mathcal{X}$ into $\mathcal{Z}$. The better a function $f(x) \coloneqq \psi(\phi(x))$ reconstructs x on a test set, the better we match the manifold assumption. On the other hand, the prototype assumption assumes that there is some set of prototypes that characterizes the healthy distribution well. In our case, the prototypes would be echos corresponding to healthy hearts, i.e., a subset of $\mathcal{X}$. Under the prototype assumption, our model f must be able to assign a given sample to one of the learned prototypes, i.e., project anomalies to the closest healthy echo.

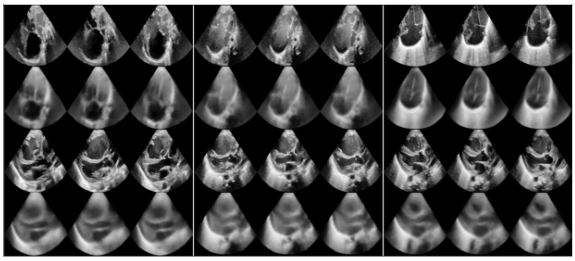
Table 1 contains the scores of the VAE, TVAE-C, TVAE-R, and TVAE-S. We report the *Mean Squared Error* (MSE), *Peak Signal to Noise Ratio* (PSNR), and *Structural Similarity Index Measure* (SSIM). We observe that TVAE-C has consistently higher MSE and SSIM errors and lower

Table 1. Apical 4-chamber view reconstruction performance on test data of the proposed approaches (TVAE-C, TVAE-R and TVAE-S) compared with the baseline (VAE). Means and standard deviations are computed across 10 validation splits.

		VAE	TVAE-C	TVAE-R	TVAE-S
SSD	MSE	**0.013**±0.0	0.014±0.0	0.014±0.0	**0.013**±0.0
	PSNR	**19.008**±0.11	18.574±0.15	18.58±0.13	18.774±0.09
	SSIM	0.545±0.01	0.544±0.01	0.545±0.01	**0.552**±0.01
RVDil	MSE	**0.012**±0.0	0.014±0.0	0.013±0.0	0.013±0.0
	PSNR	**19.146**±0.05	18.7±0.07	18.82±0.08	18.803±0.04
	SSIM	**0.555**±0.0	0.549±0.0	0.554±0.0	**0.555**±0.0
PH	MSE	**0.012**±0.0	0.014±0.0	0.014±0.0	0.014±0.0
	PSNR	**19.084**±0.07	18.66±0.07	18.723±0.08	18.727±0.07
	SSIM	0.552±0.0	0.55±0.0	0.551±0.0	**0.553**±0.0



(a) Healthy reconstructions



(b) SSD reconstructions

Figure 2. Examples of healthy (a) and SSD (b) samples (first and third rows) and their reconstructions (second and fourth rows) using the TVAE-S model. We sample 3 frames from each echo's 25 frame long sequences.

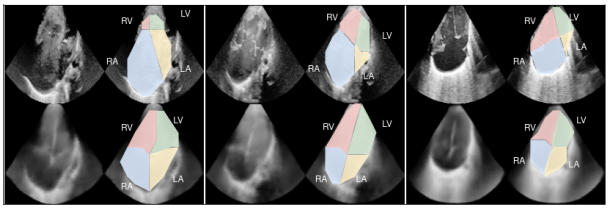


Figure 3. Projection of 4CV view anomalous echo (top) to healthy prototype (bottom). Projections of right (R) and left (L) ventricle (V) and atrium (A) are highlighted in color. The reconstruction of SSD samples approximates a healthy version of the input, e.g., by normalizing the scale of the right and left ventricles (left), adding the ventricular septum (middle), or fixing the location of the valves (right).

PSNR than both TVAE-R and TVAE-S. Upon inspection of the reconstructed test videos we notice that, for most seeds, TVAE-C converges to a local optimum where the model learns mean representations of the input videos, thus ignoring the latent dimensions containing temporal information, as described in Section 3. On the other hand, we did not observe this behavior in TVAE-R and TVAE-S, suggesting that these models indeed capture dynamic properties of echos through the learned latent representations. Additionally, TVAE-S achieves good echo reconstructions even for samples with transducer position displacement, improving upon TVAE-R and achieving similar performance as VAE despite having a smaller information bottleneck. The proposed approaches, TVAE-C, TVAE-R, and TVAE-S, encode videos into $d+2$ or $d+3$ trajectory parameters respectively, while the VAE encodes each individual frame in $\mathbb{R}^d$, resulting in a total of $25 \times d$ latent parameters. In conclusion, TVAE-S and the standard VAE fulfill the manifold assumption similarly well. Figure 2 presents reconstructed healthy and SSD samples for the 4CV and PLAX echo views.

In Figure 3, we qualitatively demonstrate that TVAE satisfies the prototype assumption. We observe how the perturbed septum and enlarged/shrunken heart chambers of SSD anomalies are projected to healthy echo reconstructions.

Appendix D provides additional reconstructions and a comprehensive performance comparison of the deterministic and variational models for the 4CV and PLAX echo views.

Table 2. Anomaly detection performance in terms of area under the curve and average precision of the proposed approaches (TVAE-C, TVAE-R, and TVAE-S) compared with the baseline (VAE) on the four-chamber view and long-axis view for the three different CHD labels. Means and standard deviations are computed on the test sets across 10 data splits. The anomalous echos are considered as the positive class. AP scores of a random classifier are 0.58 (SSD), 0.72 (RVDil), and 0.72 (PH).

		SSD		RVDil		PH	
		AUROC	AP	AUROC	AP	AUROC	AP
4CV	VAE	0.645±0.08	0.667±0.08	0.477±0.05	0.715±0.05	0.498±0.05	0.722±0.03
	TVAE-C	0.913±0.09	0.916±0.11	**0.6**±0.05	0.762±0.04	0.612±0.05	0.786±0.04
	TVAE-R	**0.917**±0.05	**0.928**±0.05	0.594±0.07	0.771±0.04	0.629±0.08	**0.797**±0.06
	TVAE-S	0.868±0.05	0.892±0.05	0.595±0.03	**0.774**±0.02	**0.649**±0.06	0.794±0.05
PLAX	VAE	0.628±0.14	0.457±0.07	0.455±0.05	0.702±0.03	0.432±0.04	0.695±0.03
	TVAE-C	0.87±0.1	0.811±0.15	0.599±0.07	**0.794**±0.04	0.631±0.05	0.818±0.03
	TVAE-R	0.877±0.08	0.826±0.1	**0.61**±0.04	**0.794**±0.02	0.629±0.06	0.817±0.03
	TVAE-S	**0.914**±0.09	**0.876**±0.14	0.592±0.05	0.791±0.03	**0.636**±0.05	**0.821**±0.02

4.2. Anomaly Detection

As described in Section 3.2.2, we detect anomalies by MAP estimation:

$$\tilde{X}_H = \arg\max_{X_H}(\log(P(Y|X_H)) + ELBO(X_H))))$$

Due to the reconstruction loss in the ELBO, this optimization problem requires us to backpropagate through the whole model in every step. As a result, inference with the standard MAP formulation is inefficient and proved infeasible for our experiments. To circumvent this problem, we assumed the reconstruction part of the ELBO to be constant and solely balanced the posterior with the KL-Divergence of the encoded $\vec{b}$, i.e., how well X_H is mapped to a standard Gaussian, thus computing

$$\tilde{X}_H = \arg\max_{X_H}(P(Y|X_H) - KL[q(\vec{b}|X_H)||p(\vec{b})])$$

Solving this optimization procedure results in only backpropagating through the encoder instead of the whole model, which leads to a significant speedup.

To optimize this objective we initialize $\tilde{X}_H$ with the reconstructions computed by the respective model, i.e. $\tilde{X}_H^{(0)} = f(Y)$ for model f and input Y. We then solve the inference problem with the Adam optimizer, incorporating a learning rate of 0.01 and taking 100 optimizer steps per sample. Additionally, we weight the TV norm with a factor of 0.001. For each sample Y, we define the anomaly score $\alpha(Y) := \frac{1}{T}\sum_t^T \|\vec{\tilde{a}}^{(j)}\|_2^2$ as described in Section 3.2.2. Anomaly detection performance is then evaluated in terms of the *Area Under the Receiver Operator Curve* (AUROC) and *Average Precision* (AP) when considering the anomalies as the positive class. In Table 2, we provide a complete overview of the results of the anomaly detection experiments over both views.

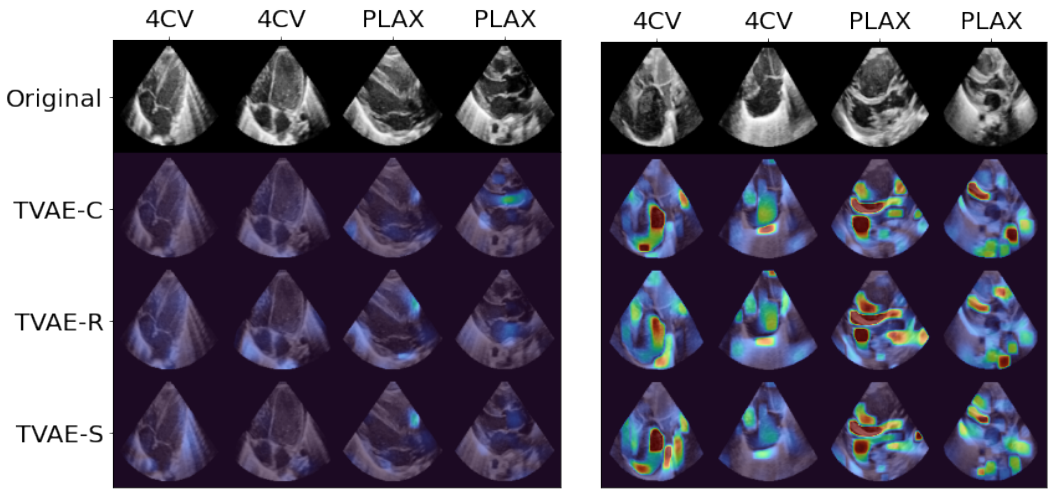
We observe that the proposed approaches outperform the VAE in all experiments. Especially when detecting SSD anomalies, our models TVAE-C, TVAE-R, and TVAE-S have significantly better performance than the standard VAE. We also note that, despite outperforming TVAE-C and TVAE-R in terms of reconstruction quality, TVAE-S does not always perform better in the anomaly detection task. We explain the score discrepancies between SSD and RVDil/PH by the fact that SSDs deviate considerably from the healthy distribution. On the contrary, RVDil and PH are more subtle and require expert knowledge and several echocardiogram views to be detected in practice.

Additionally, even though the TVAE variations have considerably fewer latent parameters ($d + 2/d + 3$) than the VAE ($25d$), they achieve similar reconstruction quality performance as demonstrated in Section 4.1. In case of VAE, this gives the optimizer more flexibility when solving the MAP problem since the frames of $\tilde{X}_H$ can be updated independently to encode them on Gaussian parameters close to $\mathcal{N}(0, \mathbb{I})$, which may result in overfitting during MAP estimation.

Another reconstruction based inference method approach where we simply define $\alpha_f(X)$ over the MSE, i.e. $\alpha_f(X) = \frac{1}{T}\sum_{j=1}^T \|(x^{(j)} - (f(X))^{(j)})\|_2^2$, is presented in Appendix E.

4.3. Decision Heatmaps

In this experiment, we present how the estimated anomaly perturbation $\tilde{A}$ can be applied to highlight anomalous regions. Intuitively, anomalous regions of input echos Y differ more substantially from its healthy projection X_H than healthy regions. Consequently, this leads to higher magnitude values in the corresponding locations in the frames of $\tilde{A}$. In turn, we are able to compute an anomaly heatmap by temporally averaging the estimated anomaly perturbation with $\frac{1}{T}\sum_{j=1}^T \vec{\tilde{a}}^{(j)}$. In Figure 4, we present examples of such maps for each TVAE variation. We observe that TVAE not only has consistently low magnitude responses for healthy



(a) Healthy (b) Anomalous

Figure 4. Anomaly response maps of TVAE-R and TVAE-S for healthy samples (a) and echos with CHDs (b). Note how healthy heatmaps are mostly constant, while anomalous maps contain regions with high responses in anomalous regions, corresponding to enlarged ventricles (first/second) or perturbed septums (third/fourth).

echos, but regions corresponding to, e.g., enlarged chambers, are well highlighted in the echos with CHDs. These heatmaps provide TVAE with an additional layer of interpretability and could foster the integration of the proposed algorithm in clinical settings, as the reason for the decisions made by TVAE can easily be interpreted by clinicians. This helps practitioners build trust in the model's decisions and provides a more intuitive explanation of its outputs. More examples of decision heatmaps are provided in Appendix F.

5. Discussion

In this work, we introduce the TVAE; a new generative model designed explicitly for echocardiogram data. We propose three variants of the model, the TVAE-C and TVAE-R, which make strong assumptions about the periodicity of the data, and the TVAE-S, which can handle more dynamic inputs. Throughout this work, we compared the proposed approach to the VAE in terms of its reconstruction performance and anomaly detection capabilities in a new in-house echo dataset consisting of two different echo views of healthy patients and patients suffering from various CHD. In exhaustive experiments, we demonstrated how TVAE can achieve reconstruction quality comparable to VAE while having a significantly smaller information bottleneck. Additionally, we verified that the proposed model can project out-of-distribution samples, i.e., patients suffering from CHD, into the subspace of healthy echos when learning normative priors and concluded that TVAE fulfills crucial assumptions for reconstruction based anomaly detection. Consequently, we evaluated CHD detection performance of our model, where we showed that it leads to a considerable improvement over frame-wise VAE with MAP-based anomaly detection. Furthermore, we demonstrated how TVAE can separate SSD anomalies almost perfectly from healthy echos. Finally, we present the ability of this model to not only detect but also localize anomalies with heatmaps generated from the MAP output, which could help clinicians with the diagnosis of CHDs.

Limitations and Future Work Even though we observe convincing results for SSD, performance for the detection of RVDil and PH is still insufficient for clinical application. This is not unexpected given that these defects are rather subtle and our relatively small in-house dataset. It would thus be interesting to apply the proposed approach to different and larger cohorts. In the future, we plan to collect more samples for our in-house dataset. With a more extensive dataset at hand, we look forward to exploring methods that would allow combinations of TVAE with one class classification or future frame prediction methods to achieve more robust anomaly detection in echocardiography-based disease detection.

The spiral trajectory of the TVAE-S model assumes continuous movement over the video and might thus still be limiting in situations where sudden movement occurs. Investigating accelerating trajectories could thus be an interesting direction. Further, we want to extend the TVAE to multiple modalities such that it is possible to train a model that learns a coherent latent trajectory of multiple echo views of the same heart. As future work, we are also interested in extending the TVAE to different types of medical modalities

by designing trajectory functions that leverage modality-specific characteristics.

References

An, J. and Cho, S. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2(1):1–18, 2015.

Baur, C., Wiestler, B., Albarqouni, S., and Navab, N. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *International MICCAI brainlesion workshop*, pp. 161–169. Springer, 2018.

Baur, C., Graf, R., Wiestler, B., Albarqouni, S., and Navab, N. Steganomaly: inhibiting cyclegan steganography for unsupervised anomaly detection in brain mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 718–727. Springer, 2020.

Baur, C., Wiestler, B., Muehlau, M., Zimmer, C., Navab, N., and Albarqouni, S. Modeling healthy anatomy with artificial intelligence for unsupervised anomaly detection in brain mri. *Radiology: Artificial Intelligence*, 3(3): e190169, 2021.

Buskens, E., Stewart, P., Hess, J., Grobbee, D., and Wladimiroff, J. Efficacy of fetal echocardiography and yield by risk category. *Obstetrics & Gynecology*, 87(3): 423–428, 1996.

Cerri, O., Nguyen, T. Q., Pierini, M., Spiropulu, M., and Vlimant, J.-R. Variational autoencoders for new physics mining at the large hadron collider. *Journal of High Energy Physics*, 2019(5):1–29, 2019.

Chalapathy, R. and Chawla, S. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

Chen, J., Sathe, S., Aggarwal, C., and Turaga, D. Outlier detection with autoencoder ensembles. In *Proceedings of the 2017 SIAM international conference on data mining*, pp. 90–98. SIAM, 2017.

Chen, X. and Konukoglu, E. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. *arXiv preprint arXiv:1806.04972*, 2018.

Chen, X., You, S., Tezcan, K. C., and Konukoglu, E. Unsupervised lesion detection via image restoration with a normative prior. *Medical image analysis*, 64:101713, 2020.

Dong, S., Luo, G., Sun, G., Wang, K., and Zhang, H. A left ventricular segmentation method on 3d echocardiography using deep learning and snake. In *2016 Computing in Cardiology Conference (CinC)*, pp. 473–476. IEEE, 2016.

Dufrenois, F. A one-class kernel fisher criterion for outlier detection. *IEEE transactions on neural networks and learning systems*, 26(5):982–994, 2014.

Gao, X., Li, W., Loomes, M., and Wang, L. A fused deep learning architecture for viewpoint classification of echocardiography. *Information Fusion*, 36:103–113, 2017.

Gautam, C., Balaji, R., Sudharsan, K., Tiwari, A., and Ahuja, K. Localized multiple kernel learning for anomaly detection: One-class classification. *Knowledge-Based Systems*, 165:241–252, 2019.

Ghafoori, Z. and Leckie, C. Deep multi-sphere support vector data description. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pp. 109–117. SIAM, 2020.

Ghasemi, A., Rabiee, H. R., Manzuri, M. T., and Rohban, M. H. A bayesian approach to the data description problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pp. 907–913, 2012.

Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., and Davis, L. S. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 733–742, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kwon, J.-m., Kim, K.-H., Jeon, K.-H., and Park, J. Deep learning for predicting in-hospital mortality among heart disease patients based on echocardiography. *Echocardiography*, 36(2):213–218, 2019.

Laumer, F., Fringeli, G., Dubatovka, A., Manduchi, L., and Buhmann, J. M. Deepheartbeat: Latent trajectory learning of cardiac cycles using cardiac ultrasounds. In *Machine Learning for Health*, pp. 194–212. PMLR, 2020.

Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E. A. R., Jodoin, P.-M., Grenier, T., et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019.

Liu, W., Luo, W., Lian, D., and Gao, S. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6536–6545, 2018.

Louis, M., Couronné, R., Koval, I., Charlier, B., and Durrleman, S. Riemannian Geometry Learning for Disease Progression Modelling. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11492 LNCS:542–553, 2019. ISSN 16113349. doi: 10.1007/978-3-030-20351-1_42.

Madani, A., Ong, J. R., Tibrewal, A., and Mofrad, M. R. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. *NPJ digital medicine*, 1(1):1–11, 2018.

Moradi, S., Oghli, M. G., Alizadehasl, A., Shiri, I., Oveisi, N., Oveisi, M., Maleki, M., and Dhooge, J. Mfp-unet: A novel deep learning based approach for left ventricle segmentation in echocardiography. *Physica Medica*, 67: 58–69, 2019.

Moya, M. M. and Hush, D. R. Network constraints and multi-objective optimization for one-class classification. *Neural networks*, 9(3):463–474, 1996.

Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C. P., Heidenreich, P. A., Harrington, R. A., Liang, D. H., Ashley, E. A., et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580 (7802):252–256, 2020.

Pang, G., Shen, C., Cao, L., and Hengel, A. V. D. Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, 54(2):1–38, 2021.

Park, D., Hoshi, Y., and Kemp, C. C. A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder. *IEEE Robotics and Automation Letters*, 3(3):1544–1551, 2018.

Pawlowski, N., Lee, M. C., Rajchl, M., McDonagh, S., Ferrante, E., Kamnitsas, K., Cooke, S., Stevenson, S., Khetani, A., Newman, T., et al. Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders. 2018.

Pinaya, W. H. L., Tudosiu, P.-D., Gray, R., Rees, G., Nachev, P., Ourselin, S., and Cardoso, M. J. Unsupervised brain anomaly detection and segmentation with transformers. *arXiv preprint arXiv:2102.11650*, 2021.

Principi, E., Vesperini, F., Squartini, S., and Piazza, F. Acoustic novelty detection with adversarial autoencoders. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 3324–3330. IEEE, 2017.

Ratsch, G., Mika, S., Scholkopf, B., and Muller, K.-R. Constructing boosting algorithms from svms: An application to one-class classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1184–1199, 2002.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. Deep one-class classification. In *International conference on machine learning*, pp. 4393–4402. PMLR, 2018.

Ruff, L., Vandermeulen, R. A., Franks, B. J., Müller, K.-R., and Kloft, M. Rethinking assumptions in deep anomaly detection. *arXiv preprint arXiv:2006.00339*, 2020.

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G., and Müller, K.-R. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 2021.

Sabokrou, M., Khalooei, M., Fathy, M., and Adeli, E. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3379–3388, 2018.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7): 1443–1471, 2001.

Singh, Y. and McGeoch, L. Fetal anomaly screening for detection of congenital heart defects. *J. Neonatal Biol.*, 5 (2):100–115, 2016.

Tax, D. M. and Duin, R. P. Support vector data description. *Machine learning*, 54(1):45–66, 2004.

Van Der Linde, D., Konings, E. E., Slager, M. A., Witsenburg, M., Helbing, W. A., Takkenberg, J. J., and Roos-Hesselink, J. W. Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. *Journal of the American College of Cardiology*, 58(21): 2241–2247, 2011.

Van Velzen, C., Clur, S., Rijlaarsdam, M., Pajkrt, E., Bax, C., Hruda, J., de Groot, C., Blom, N., and Haak, M. Prenatal diagnosis of congenital heart defects: accuracy and discrepancies in a multicenter cohort. *Ultrasound in Obstetrics & Gynecology*, 47(5):616–622, 2016.

Vaseli, H., Liao, Z., Abdi, A. H., Girgis, H., Behnami, D., Luong, C., Dezaki, F. T., Dhungel, N., Rohling, R., Gin, K., et al. Designing lightweight deep learning models for echocardiography view classification. In *Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 10951, pp. 93–99. SPIE, 2019.

Xu, D., Ricci, E., Yan, Y., Song, J., and Sebe, N. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*, 2015.

Xu, H., Chen, W., Zhao, N., Li, Z., Bu, J., Li, Z., Liu, Y., Zhao, Y., Pei, D., Feng, Y., et al. Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. In *Proceedings of the 2018 world wide web conference*, pp. 187–196, 2018.

Yan, S., Smith, J. S., Lu, W., and Zhang, B. Abnormal event detection from videos using a two-stream recurrent variational autoencoder. *IEEE Transactions on Cognitive and Developmental Systems*, 12(1):30–42, 2018.

You, S., Tezcan, K. C., Chen, X., and Konukoglu, E. Unsupervised lesion detection via image restoration with a normative prior. In *International Conference on Medical Imaging with Deep Learning*, pp. 540–556. PMLR, 2019.

Yu, G., Wang, S., Cai, Z., Zhu, E., Xu, C., Yin, J., and Kloft, M. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 583–591, 2020.

Zeiler, M. D., Krishnan, D., Taylor, G. W., and Fergus, R. Deconvolutional networks. In *2010 IEEE Computer Society Conference on computer vision and pattern recognition*, pp. 2528–2535. IEEE, 2010.

A. Variational Trajectory Model ELBO derivation

Recall that we define $\vec{b} \sim q_\theta(\vec{b}|X) \coloneqq \mathcal{N}(\vec{\mu_b}, diag(\vec{\sigma_b}))$ with prior $p(\vec{b}) \coloneqq \mathcal{N}(0, \mathbb{I})$, while leaving the other trajectory parameters deterministic. Note that this effectively means that we define uniform priors $p(f)$, $p(\omega)$ and $p(v)$ over their support, while having posteriors

$$q_\theta(f|X) \coloneqq \delta_{\phi_f(X)}(f), \quad q_\theta(\omega|X) \coloneqq \delta_{\phi_\omega(X)}(\omega), \quad q_\theta(v|X) \coloneqq \delta_{\phi_v(X)}(v)$$

where δ_y is the Dirac Delta spiking at y and $\phi_f(X)$, $\phi_\omega(X)$ and $\phi_v(X)$ are the trajectory parameter outputs of the encoder ϕ with weights θ for f, ω and v respectively.

Given input sample x and latent z, recall that VAEs aim to maximize the *Evidence LOwer Bound* (ELBO):

$$E_{q_\theta(z|x)}[\log(p_\eta(x|z))] - KL[q_\theta(z|x)||p(z)]$$

Here, x corresponds to the input echocardiogram $X \coloneqq (\vec{x^{(j)}}, t^{(j)})_{j=1}^T$ whereas $z \coloneqq (\vec{b}, f, \omega, v)$.

Note that $\vec{b}, f, \omega$ and v are conditionally independent, i.e.

$$q_\theta(\vec{b}, f, \omega, v|X) = q_\theta(\vec{b}|X) q_\theta(f|x) q_\theta(\omega|X) q_\theta(v|X)$$

The KL divergence is additive for joint distributions of independent random variables, i.e. for $P = (P_1, P_2)$ and $Q = (Q_1, Q_2)$, where P_1, P_2, Q_1 and Q_2 are independent, it holds that

$$KL(P||Q) = KL(P_1||Q_1) + KL(P_2||Q_2)$$

We can thus rewrite the ELBO as

$$\begin{aligned} &E_{q_\theta(\vec{b}, f, \omega, v|X)}[\log(p_\eta(X|\vec{b}, f, \omega, v))] \\ -&KL[q_\theta(\vec{b}|X)||p(\vec{b})] - KL[q_\theta(f|X)||p(f)] \\ -&KL[q_\theta(\omega|X)||p(\omega)] - KL[q_\theta(v|X)||p(v)] \end{aligned}$$

Since we assumed a uniform prior for f, ω and v, their KL-Divergence terms become constant under the Dirac Delta distribution. We can thus ignore the respective terms in the ELBO during optimization as they do not change the result of the $argmax$.

Additionally, since

$$\int \delta_y(x) f(x) dx = f(y)$$

we can rewrite the ELBOs reconstruction term as

$$\begin{aligned} &E_{q_\theta(\vec{b}, f, \omega, v|X)}[\log(p_\eta(X|\vec{b}, f, \omega, v))] \\ &\quad = \int \delta_{\phi_f(X)}(f) \delta_{\phi_\omega(X)}(\omega) \delta_{\phi_v(X)}(v) q_\theta(\vec{b}|X) \log(p_\eta(X|\vec{b}, f, \omega, v)) d\vec{b} df d\omega dv \\ &\quad = \int q_\theta(\vec{b}|X) \log(p_\eta(X|\vec{b}, \phi_f(X), \phi_\omega(X), \phi_v(X))) d\vec{b} \\ &\quad = E_{q_\theta(\vec{b}|X)}[\log(p_\eta(X|\vec{b}, \phi_f(X), \phi_\omega(X), \phi_v(X)))] \end{aligned}$$

Finally, this leads to the following reformulation of the ELBO objective:

$$E_{q_\theta(\vec{b}|X)}[\log(p_\eta(X|\vec{b}, \phi_f(X), \phi_\omega(X), \phi_v(X)))] - KL[q_\theta(\vec{b}|X)||p(\vec{b})]$$

B. Cohort Examples

The dataset for this study consists of echos of 192 newborns and infants up to one year of age collected between 2019 and 2020 at a single center by a single pediatric cardiologist. All examinations were performed with the GE Logic S8

Table 3. Cohort Statistics

Feature	**Statistic**
No. of Patients	192
No. of Patients with no CHD	69
No. of Patients with SSD	5
No. of Patients with PH	73
No. of Patients with RVDIL	73
Age (Days) (Mean± SD)	34 ± 48
Time until birth (Days) (Mean± SD)	232 ± 46
Weight (Gramms) (Mean± SD)	2774 ± 1227
Manufacturer (Ultrasound Machine / Transducer)	GE Logic S8 / S4-10 at 6 MHz
Original Video Size (pixel×pixels)	1440×866
Video length (frames) (Mean± SD)	122 ± 7
Video FPS	25 fps

ultrasound machine and contain 2D video sequences of at least 2 standard echo views, i.e., apical 4-chamber view (4CV) and parasternal long-axis view (PLAX). Of the 192 patients, 123 suffer from, potentially multiple, CHDs, and 69 are healthy. See Table 3 for more details.

In order to evaluate anomaly detection performance, the dataset was labeled in three different categories by a pediatric cardiologist. These include *Pulmonary Hypertension* (PH), *Right Ventricular Dilation* (RVDil) and *Severe Structural Defects* (SSD). While PH and RVDil are well-defined pathologies, SSD was defined as a category of multiple rare but severe CHD pathologies, including Ebstein's anomaly, anomalous left coronary artery origin from pulmonary artery (ALCAPA), atrio-ventricular discordance, and ventricular-artery concordance (AVD-VAC), Shone-complex, total anomalous pulmonary venous drainage (TAPVD), tetralogy of fallot (ToF) and complete atrioventricular septal defect (cAVSD). We illustrate examples for healthy, SSD, PH, and RVDil echos of both 4CV and PLAX views in Figure 5.

The collected echocardiograms were preprocessed by resizing them to 128×128 pixels. Additionally, histogram equalization was performed to increase the contrast of the frames, and pixel values were normalized to the range $[0, 1]$. For video inputs, we assume that any heart anomaly should always be visible for a certain period over the heart cycle. It thus suffices to have a model that reconstructs only a fixed number of video frames, as long as at least one heart cycle is present in the video. The collected videos are recorded with 24 frames per second (FPS), and we assume that a heart beats at least 30 times a minute. Therefore, we decided to subsample the video frequency to 12 FPS and reconstruct videos with a fixed length of 25 frames, which is enough to capture at least one cycle in every video. Hence, the input for video models consists of 25 concatenated consecutive frames of the subsampled video. Having fixed length inputs enables us to implement more efficient architectures.

As in most clinical applications, the scarcity of the data often leads to overfitting. To prevent this, we apply data augmentation during training by transforming samples with random affine transformations, brightness adjustments, gamma corrections, blurring, and the addition of Salt and Pepper noise before performing the forward pass.

C. Architecture

As described in Appendix B, video inputs consist of 25 concatenated frames $(x^{(i)}, t^{(i)})$ with timestamps $t^{(i)}$. Hence, we can treat video frames like different channels of an image, and pass them to a *residual* (He et al., 2016) encoder backbone. Each frame $(x^{(i)}, t^{(i)})$ is then individually decoded by passing $\vec{\ell}_{circular}(t^{(i)})$, $\vec{\ell}_{rot}(t^{(i)})$ or $\vec{\ell}_{spiral}(t^{(i)})$ to a *deconvolution* (Zeiler et al., 2010) based decoder. To train the VAE, we used identical encoder and decoder architectures, only changing the first layer to take a single grayscale channel instead of 25 frames and adapting latent fully connected layers to match dimensions. We provide schematics for the building blocks of our architectures in Figure 6 and describe the encoder/decoder architecture of our experiments in Figure 7.

Table 4 contains the hyperparameters used in our experiments. Except for the number of steps, we kept hyperparameters mostly the same for all models. This is because, in contrast to the frame-wise models, TAE and TVAE models required more steps to converge. We suspect this is because the dimensionality of the input is 25 times larger, and the model thus

Table 4. Hyperparameters chosen across our experiments.

Hyperparameter	AE/VAE	TAE/TVAE
Latent Dimension	64	66/67 (b:64; f:1; ω:1; v:1)
Batch Size	128	64
Steps	5000	106500
Number of Frames	1	25
Optimizer	Adam	Adam
Learning Rate	10^{-4}	10^{-4}
Reconstruction Loss	MSE	MSE
VAE β	1	1

requires more parameter updates to converge to a suitable optima that results in good reconstructions. Batch size was chosen according to GPU memory capacity. All models are pretrained on the EchoDynamic dataset to speed up training convergence (Ouyang et al., 2020).

D. Further Reconstruction Experiments

In addition to the reconstruction quality experiments provided in Section 4.1, we compared the performance of the variational models to deterministic ones (i.e., standard autoencoder and non-variational trajectory models). As seen in Table 5, the deterministic trajectory models result in a similar performance to the variational models and are even slightly better with respect to the structural similarity score. Even though trained on the same architecture and for the same number of steps as the VAE, the autoencoder did not produce very good reconstruction scores in this experiment. We suspect that this may be an artifact of overfitting due to the small training set.

We provide more reconstructions of TVAE-S in Figure 8.

E. Reconstruction error based anomaly detection

A common alternative to MAP-based anomaly detection is the detection of anomalies purely based on the reconstruction error of the model. This means, for model f, sample $x \in \mathcal{X}$ and data space $\mathcal{X}$, we would simply define $\alpha_f(x) = \|x - f(x)\|_2^2$. In order to quantify the performance of non-variational dynamic trajectory model (TAE) and compare to a standard autoencoder trained on single frame reconstruction, we performed another ablation on AE, VAE, and the variants of TAE and TVAE. Results of this ablation are aggregated in Table 6.

F. More Decision Heatmaps

In addition to the heatmaps presented in Section 4.3, we provide a more extensive collection of TVAE-S decision heatmaps in Figure 9 and Figure 10.

G. Generated Videos

The introduced TVAE variations are generative models. As such, in addition to producing good reconstructions of existing samples, they allow us to sample from the learned distribution. To qualitatively validate generative performance, we provide random generations of the TVAE-S model in Figure 11 for both 4CV and PLAX views.

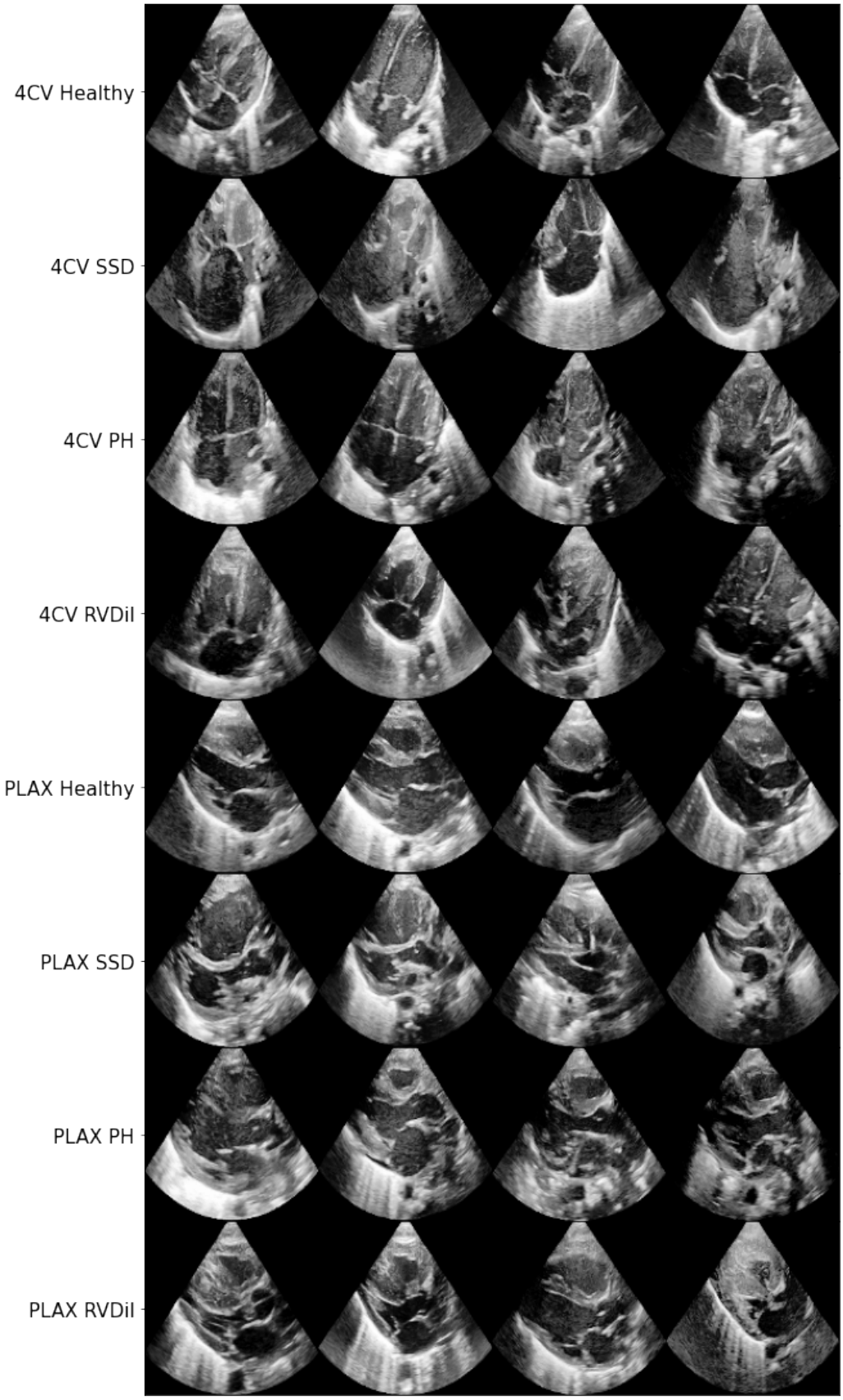
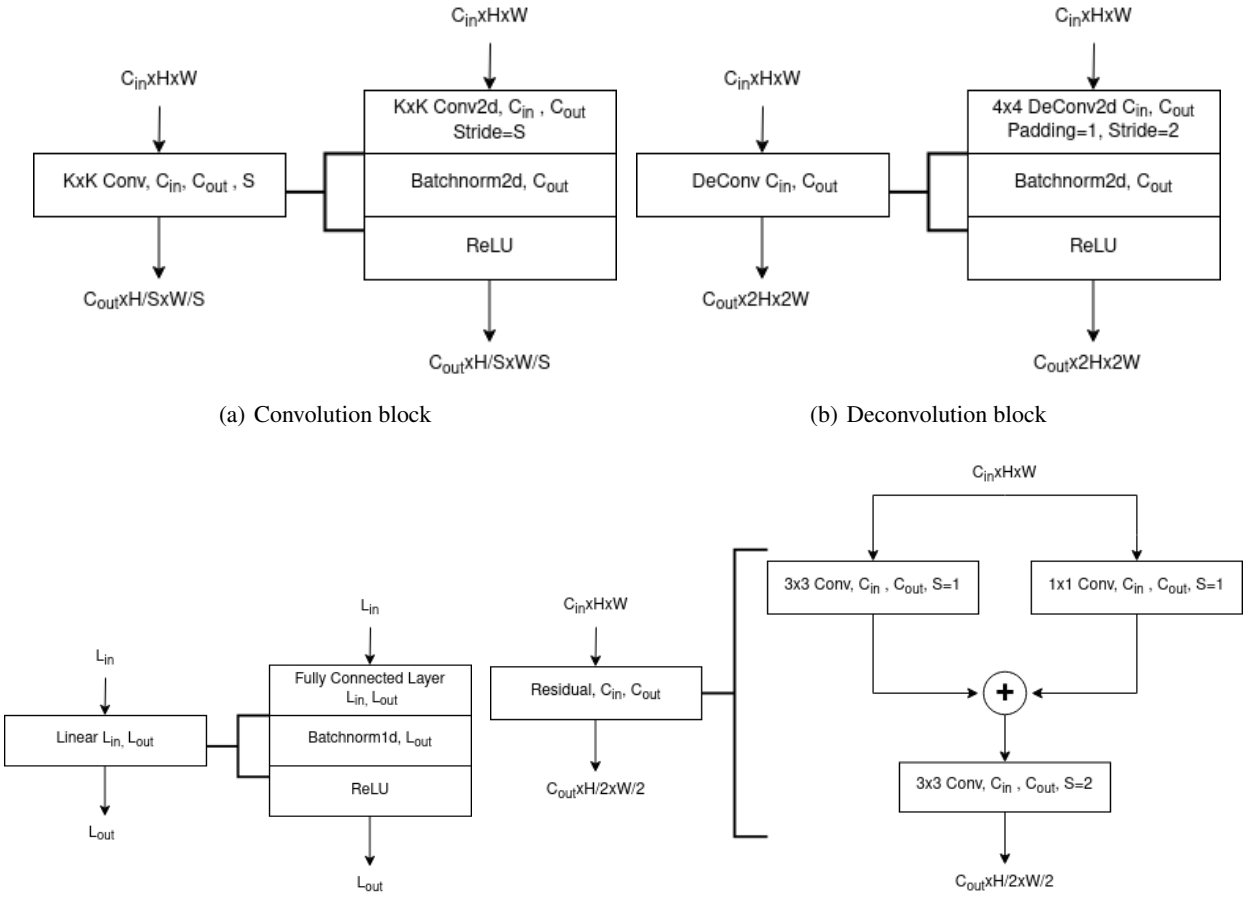


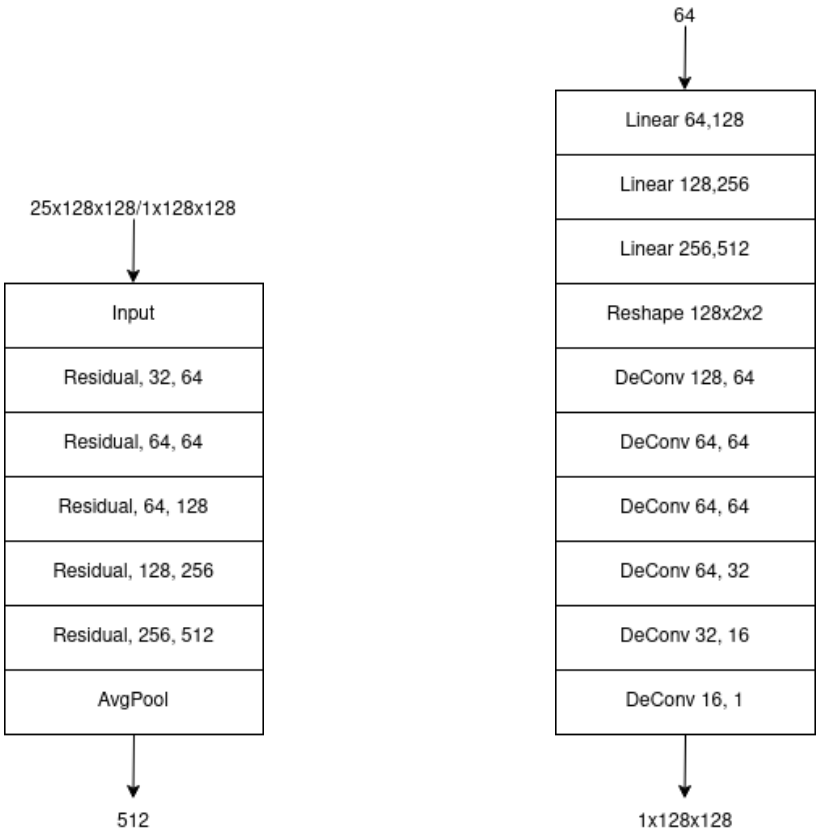
Figure 5. Examples of each label of the cohort in 4CV and PLAX views.



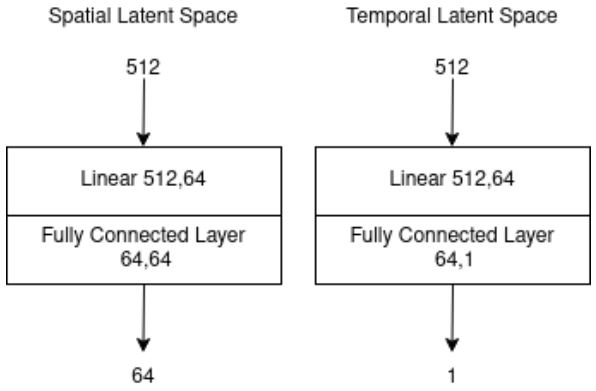
(a) Convolution block (b) Deconvolution block

(c) Linear block (d) Residual block

Figure 6. Definitions of the encoder/decoder building blocks.



(a) Encoder (b) Decoder



(c) Latent space components

Figure 7. Architectures of encoder (a), decoder (b) and latent space components (c). Spatial latent space components are used to learn z, μ and σ for AE/VAE or b, μ_b and σ_b for TAE/TVAE. Temporal latent space components learn f, ω or v for TAE/TVAE.

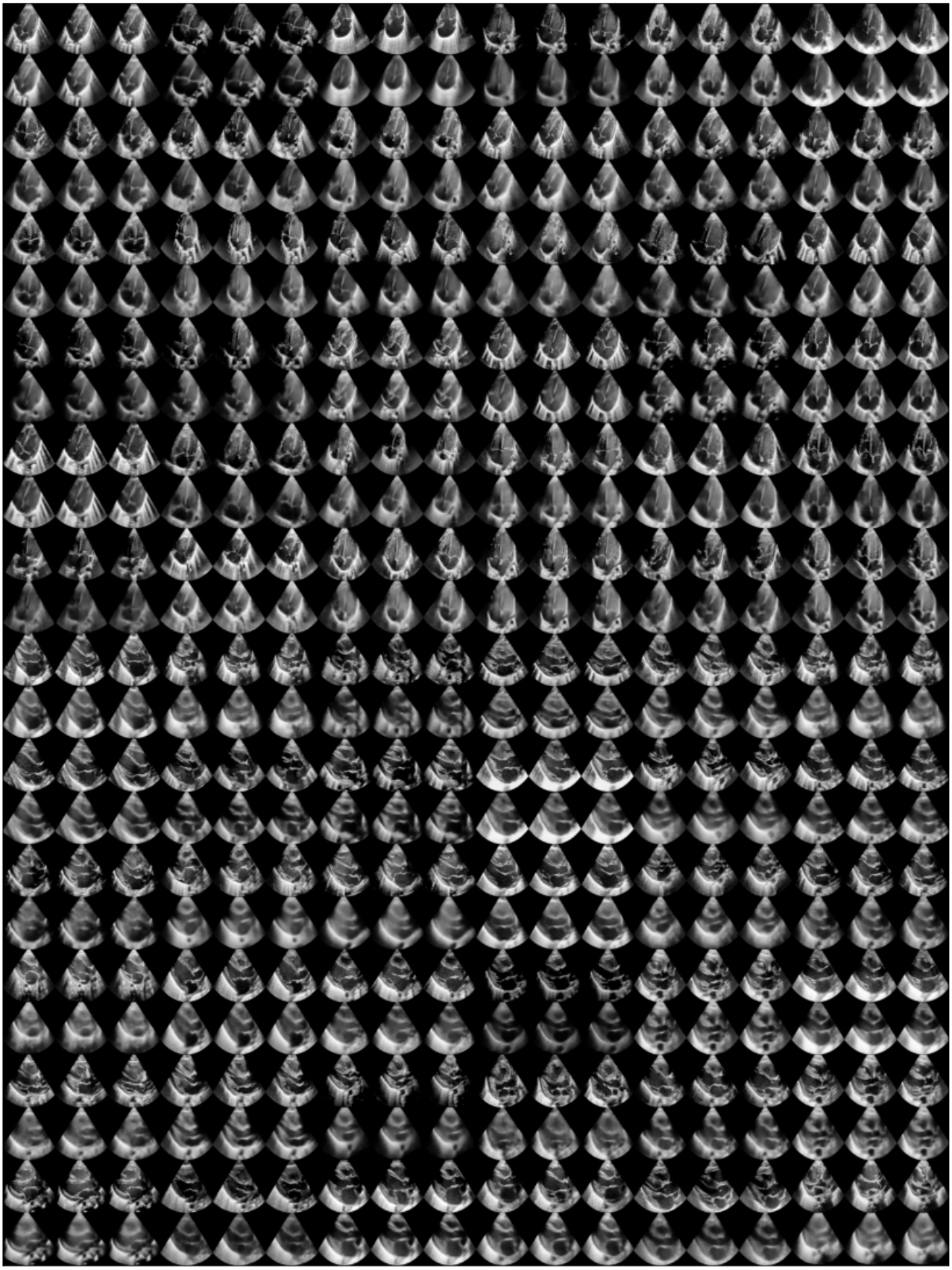


Figure 8. More TVAE-S reconstructions.

Table 5. Reconstruction scores across all introduced models and compared to AE/VAE.

		SSD			RVDil			PH		
		MSE	PSNR	SSIM	MSE	PSNR	SSIM	MSE	PSNR	SSIM
4CV	AE	0.427±0.07	4.692±0.65	0.246±0.01	0.526±0.1	3.722±0.78	0.246±0.0	0.57±0.08	3.248±0.54	0.244±0.0
	VAE	**0.013**±0.0	**19.02**±0.11	0.546±0.01	**0.012**±0.0	**19.146**±0.05	0.555±0.0	**0.012**±0.0	**19.093**±0.07	0.553±0.0
	TAE-C	**0.013**±0.0	18.922±0.13	0.553±0.01	0.013±0.0	18.933±0.06	0.557±0.0	0.013±0.0	18.905±0.04	0.556±0.0
	TAE-R	**0.013**±0.0	18.964±0.15	0.555±0.01	0.013±0.0	19.065±0.05	**0.563**±0.0	0.013±0.0	19.017±0.07	**0.562**±0.0
	TAE-S	**0.013**±0.0	18.949±0.1	**0.557**±0.01	0.013±0.0	18.926±0.06	0.559±0.0	0.013±0.0	18.897±0.07	0.558±0.0
	TVAE-C	0.014±0.0	18.562±0.15	0.544±0.01	0.014±0.0	18.707±0.07	0.549±0.0	0.014±0.0	18.661±0.07	0.55±0.0
	TVAE-R	0.014±0.0	18.568±0.15	0.545±0.01	0.013±0.0	18.813±0.09	0.554±0.0	0.014±0.0	18.73±0.08	0.552±0.0
	TVAE-S	**0.013**±0.0	18.784±0.09	0.551±0.01	0.013±0.0	18.797±0.04	0.554±0.0	0.013±0.0	18.747±0.08	0.554±0.0
PLAX	AE	0.518±0.14	3.724±1.01	0.226±0.0	0.649±0.07	2.545±0.5	0.23±0.0	0.748±0.11	1.935±0.59	0.228±0.0
	VAE	**0.017**±0.0	**17.874**±0.15	0.524±0.0	**0.015**±0.0	**18.394**±0.05	**0.541**±0.0	**0.015**±0.0	**18.442**±0.03	**0.542**±0.0
	TAE-C	0.018±0.0	17.576±0.12	0.518±0.0	0.016±0.0	17.986±0.1	0.532±0.0	0.016±0.0	18.117±0.07	0.536±0.0
	TAE-R	**0.017**±0.0	17.836±0.15	**0.528**±0.01	**0.015**±0.0	18.196±0.13	0.54±0.0	**0.015**±0.0	18.192±0.12	0.54±0.0
	TAE-S	**0.017**±0.0	17.785±0.12	0.525±0.0	**0.015**±0.0	18.184±0.04	0.539±0.0	**0.015**±0.0	18.26±0.04	**0.542**±0.0
	TVAE-C	0.019±0.0	17.264±0.15	0.509±0.01	0.017±0.0	17.795±0.05	0.526±0.0	0.017±0.0	17.778±0.07	0.526±0.0
	TVAE-R	0.018±0.0	17.555±0.18	0.519±0.01	0.016±0.0	17.958±0.08	0.533±0.0	0.016±0.0	17.971±0.12	0.533±0.0
	TVAE-S	0.018±0.0	17.633±0.09	0.517±0.0	0.016±0.0	18.115±0.1	0.536±0.0	**0.015**±0.0	18.152±0.08	0.537±0.0

Table 6. Area under the curve and average precision for experiments performed with anomaly score $\alpha_f(x) = \frac{1}{T}\sum_{t=1}^{T} \|x^{(t)} - f^{(t)}(x)\|_2^2$.

		SSD		RVDil		PH	
		AUROC	AP	AUROC	AP	AUROC	AP
4CV	AE	$0.566_{\pm 0.1}$	$0.602_{\pm 0.09}$	$\mathbf{0.634}_{\pm 0.05}$	$\mathbf{0.816}_{\pm 0.03}$	$0.612_{\pm 0.03}$	$0.803_{\pm 0.02}$
	VAE	$\mathbf{0.699}_{\pm 0.09}$	$0.732_{\pm 0.08}$	$0.619_{\pm 0.05}$	$0.803_{\pm 0.03}$	$\mathbf{0.635}_{\pm 0.04}$	$\mathbf{0.808}_{\pm 0.04}$
	TAE-C	$0.572_{\pm 0.09}$	$0.651_{\pm 0.05}$	$0.581_{\pm 0.04}$	$0.775_{\pm 0.03}$	$0.609_{\pm 0.02}$	$0.795_{\pm 0.01}$
	TAE-R	$0.612_{\pm 0.06}$	$0.695_{\pm 0.06}$	$0.594_{\pm 0.04}$	$0.781_{\pm 0.03}$	$0.617_{\pm 0.02}$	$0.8_{\pm 0.02}$
	TAE-S	$0.558_{\pm 0.1}$	$0.646_{\pm 0.08}$	$0.612_{\pm 0.04}$	$0.794_{\pm 0.03}$	$0.614_{\pm 0.03}$	$0.802_{\pm 0.02}$
	TVAE-C	$0.672_{\pm 0.06}$	$0.736_{\pm 0.05}$	$0.6_{\pm 0.04}$	$0.779_{\pm 0.03}$	$0.622_{\pm 0.03}$	$0.803_{\pm 0.01}$
	TVAE-R	$0.673_{\pm 0.07}$	$\mathbf{0.745}_{\pm 0.07}$	$0.611_{\pm 0.03}$	$0.787_{\pm 0.02}$	$0.621_{\pm 0.04}$	$0.803_{\pm 0.03}$
	TVAE-S	$0.616_{\pm 0.1}$	$0.679_{\pm 0.05}$	$0.61_{\pm 0.05}$	$0.786_{\pm 0.04}$	$0.631_{\pm 0.03}$	$0.8_{\pm 0.03}$
PLAX	AE	$0.917_{\pm 0.08}$	$0.889_{\pm 0.1}$	$0.681_{\pm 0.03}$	$0.848_{\pm 0.02}$	$0.637_{\pm 0.05}$	$\mathbf{0.827}_{\pm 0.03}$
	VAE	$0.918_{\pm 0.07}$	$0.915_{\pm 0.05}$	$\mathbf{0.683}_{\pm 0.03}$	$\mathbf{0.851}_{\pm 0.02}$	$0.631_{\pm 0.05}$	$0.819_{\pm 0.03}$
	TAE-C	$0.917_{\pm 0.09}$	$0.885_{\pm 0.13}$	$0.65_{\pm 0.03}$	$0.825_{\pm 0.02}$	$\mathbf{0.646}_{\pm 0.06}$	$\mathbf{0.827}_{\pm 0.03}$
	TAE-R	$0.913_{\pm 0.08}$	$0.878_{\pm 0.1}$	$0.68_{\pm 0.05}$	$0.84_{\pm 0.03}$	$0.639_{\pm 0.05}$	$0.823_{\pm 0.02}$
	TAE-S	$0.927_{\pm 0.07}$	$0.905_{\pm 0.09}$	$0.666_{\pm 0.03}$	$0.836_{\pm 0.01}$	$0.635_{\pm 0.04}$	$\mathbf{0.827}_{\pm 0.02}$
	TVAE-C	$0.927_{\pm 0.07}$	$0.907_{\pm 0.08}$	$0.65_{\pm 0.04}$	$0.828_{\pm 0.02}$	$0.617_{\pm 0.05}$	$0.814_{\pm 0.03}$
	TVAE-R	$0.917_{\pm 0.07}$	$0.883_{\pm 0.11}$	$0.664_{\pm 0.05}$	$0.831_{\pm 0.03}$	$0.622_{\pm 0.05}$	$0.818_{\pm 0.02}$
	TVAE-S	$\mathbf{0.935}_{\pm 0.08}$	$\mathbf{0.916}_{\pm 0.11}$	$0.658_{\pm 0.05}$	$0.828_{\pm 0.02}$	$0.625_{\pm 0.04}$	$0.823_{\pm 0.02}$

Figure 9. More TVAE-S decision heatmaps for healthy echos.

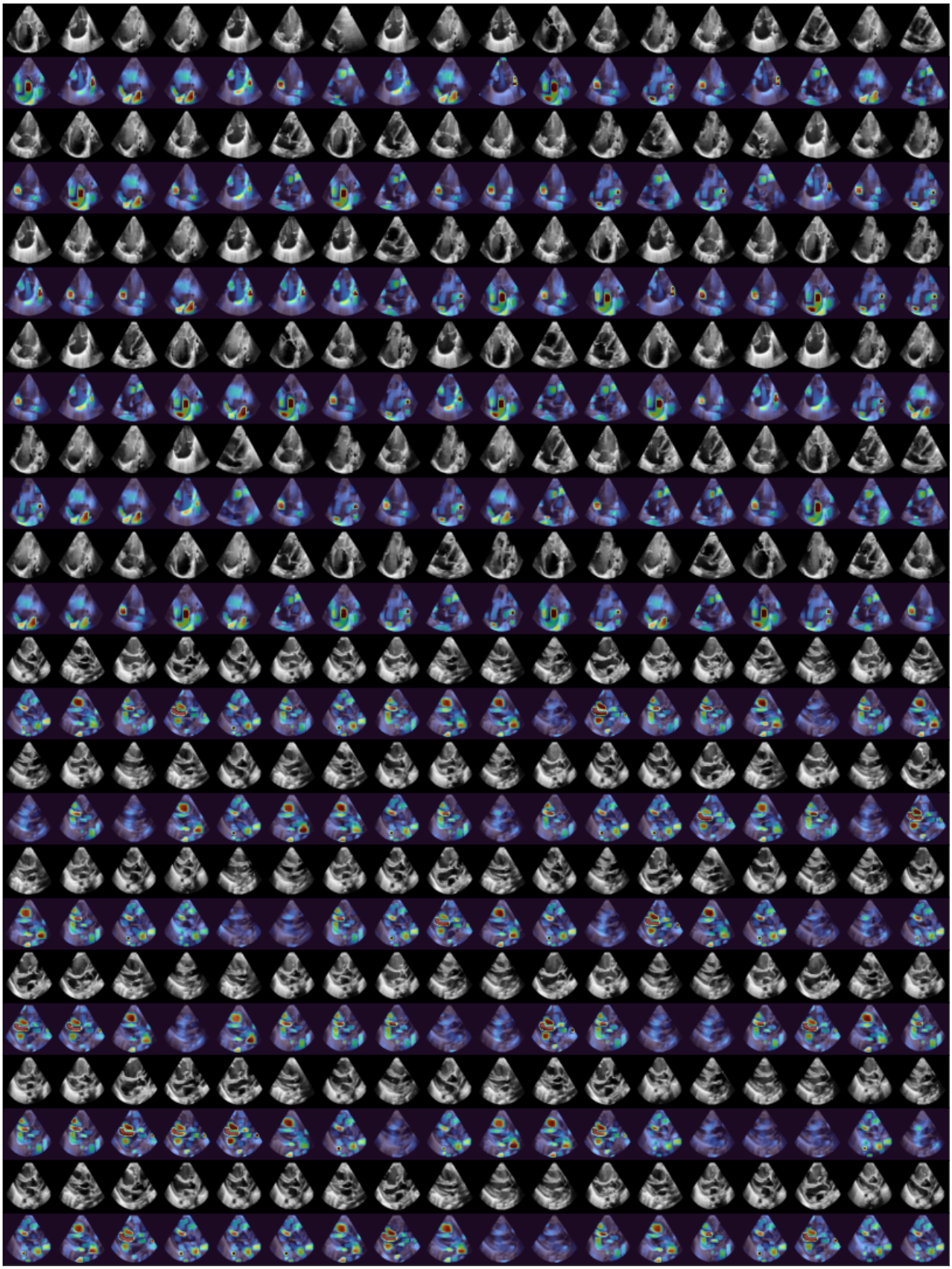


Figure 10. More TVAE-S decision heatmaps for anomalous echos.

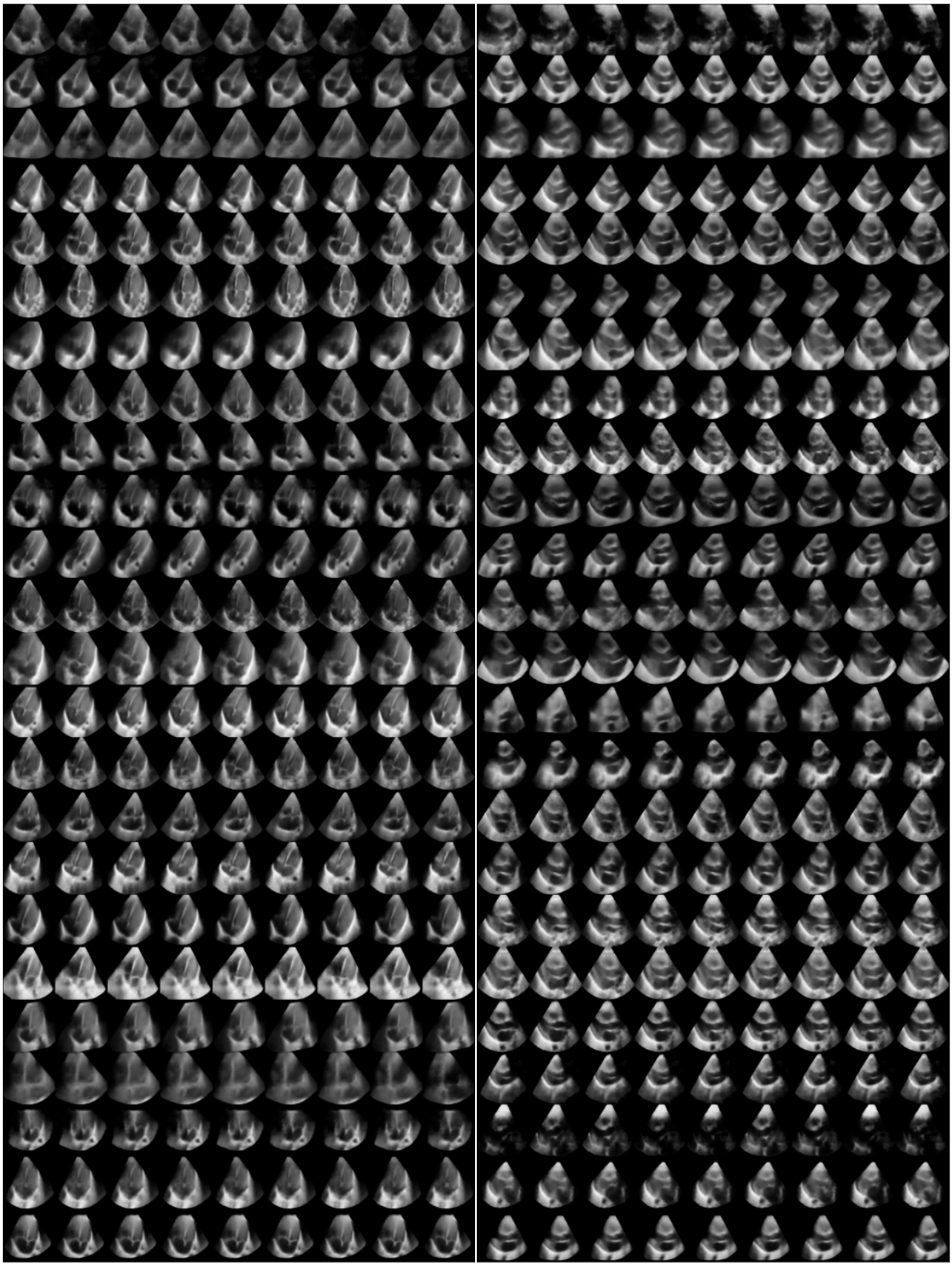


Figure 11. Random TVAE-S generations of samples in 4CV and PLAX views.