

CardiacNet: Learning to Reconstruct Abnormalities for Cardiac Disease Assessment from Echocardiogram Videos

Jiewen Yang¹, Yiqun Lin¹, Bin Pu¹, Jiarong Guo¹,
Xiaowei Xu³, and Xiaomeng Li^{1,2}

¹ The Hong Kong University of Science and Technology

{jyangcu, ylindw, jguoaz}@connect.ust.hk, {eebinpu, eexmli[✉]}@ust.hk

² HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian,
Shenzhen, China

³ Guangdong Provincial People’s Hospital, Guangzhou, China
xiao.wei.xu@foxmail.com[✉]

Abstract. Echocardiogram video plays a crucial role in analysing cardiac function and diagnosing cardiac diseases. Current deep neural network methods primarily aim to enhance diagnosis accuracy by incorporating prior knowledge, such as segmenting cardiac structures or lesions annotated by human experts. However, diagnosing the inconsistent behaviours of the heart, which exist across both spatial and temporal dimensions, remains extremely challenging. For instance, the analysis of cardiac motion acquires both spatial and temporal information from the heartbeat cycle. To address this issue, we propose a novel reconstruction-based approach named **CardiacNet** to learn a better representation of local cardiac structures and motion abnormalities through echocardiogram videos. CardiacNet accompanied by the **Consistency Deformation Codebook (CDC)** and the **Consistency Deformed-Discriminator (CDD)** to learn the commonalities across abnormal and normal samples by incorporating cardiac prior knowledge. In addition, we propose benchmark datasets named **CardiacNet-PAH** and **CardiacNet-ASD** for evaluating the effectiveness of cardiac disease assessment. In experiments, our CardiacNet can achieve state-of-the-art results in three different cardiac disease assessment tasks on public datasets CAMUS, EchoNet, and our datasets. The code and dataset are available at: <https://github.com/xmedlab/CardiacNet>

1 Introduction

Echocardiogram video, being the most widely used and easily accessible imaging modality in the field of cardiac medicine, has been proposed as a valuable tool for assessing various cardiac diseases, such as congenital heart disease [12, 22] and atypical cardiac motion [27, 36, 43]. Currently, there are several artificial intelligence methods [5, 18, 19, 25, 28–32, 34, 41, 42, 47, 49, 51] available for the assessment and evaluation of cardiac conditions in echocardiography. For instance, EchoNet [25], a state-of-the-art cardiac assessment method, employs an R2+1D network to extract global spatiotemporal features from echocardiogram videos

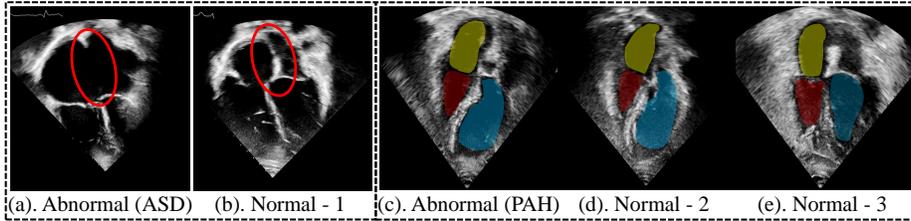


Fig. 1: Five examples from CardiacNet-PAH and CardiacNet-ASD datasets. The appearance between the Atrial Septal Defect (ASD) (a) and normal (b) is easy to distinguish. For (c), (d) and (e), the appearance of cardiac structures in the Pulmonary Arterial Hypertension patient (c) and normal (d) are similar. In contrast, normal cases (d) and (e) show significant differences. Indicates that using a single image is not able to diagnose this type of cardiac disease. Clinically, experienced physicians will use echocardiogram videos with cardiac motion information to make diagnoses.

for predicting ejection fraction (EF). However, while these methods excel at capturing spatiotemporal information, they tend to neglect the local characteristics of cardiac structure, specifically the cyclical heartbeat motion. Furthermore, their performance is still limited, which restricts their adaptability to a broader range of cardiac diseases.

To develop a general approach for cardiac disease assessment, we have identified two important characteristics that encompass a wide range of common cardiac conditions, including EF, Pulmonary Arterial Hypertension (PAH), and Atrial Septal Defect (ASD). Specifically, **(1) Local Structure Abnormality** refers to cardiac diseases that exhibit clear and distinctive abnormalities in a localized region within a single frame of an echocardiogram video. As depicted in Fig. 1(a-b), a hole (highlighted in Red) can be observed in the atrial septum, enabling the mixing of blood between the left and right atria. **(2) Cardiac Motion Abnormality** refers to cardiac diseases that may not have clear distinctive abnormalities in a single frame of echocardiogram videos, but can be detected through motion abnormalities of local cardiac structure observed in videos. For instance, in Fig. 1(c-d), there are no clear differences in cardiac structures between PAH patients and normal individuals based on a single frame of echocardiogram videos. Therefore, it is highly necessary to develop an approach to learn a better representation across both temporal and spatial patterns of local cardiac structures via echocardiography.

Existing classification and regression-based disease assessment approaches [9, 20, 25, 48] typically focus on global information and show difficulty in capturing local representations. In contrast, the reconstruction-based approaches [10, 11, 33, 34, 40] offer a more intuitive solution by accurately reconstructing the abnormal and normal cases, enabling a deeper understanding of abnormality distribution, capturing fine-grained details, and achieving accurate disease assessment results. However, existing reconstruction-based approaches were mainly designed for computed tomography (CT), magnetic resonance imaging (MRI), and X-ray modalities, focusing on abnormalities with low-level details such as tumors, bone fractures, and anomalous cardiac structures [34]. When directly applying these approaches to our datasets, their performance in assessing specific cardiac dis-

Table 1: Summary statistics of datasets CardiacNet-PAH and CardiacNet-ASD and two public datasets CAMUS [13] and EchoNet [25].

Dataset	CardiacNet-PAH (Ours)						CardiacNet-ASD (Ours)					
Attributes	Total Videos	Total Images	PAH Cases	Normal Cases	Other Cases	Resolution	Total Videos	Total Images	ASD Cases	Normal Cases	Other Cases	Resolution
	496	44,363	342	154	0	720p	231	13,471	100	131	0	720p
Dataset	CAMUS [13]						EchoNet-Dynamic [25]					
Attributes	Total Videos	Total Images	EF \geq 55% Cases	EF \leq 50% Cases	50% $<$ EF $<$ 55% Cases	Resolution	Total Videos	Total Images	EF \geq 55% Cases	EF \leq 50% Cases	50% $<$ EF $<$ 55% Cases	Resolution
	500	10,000	201	178	121	480p	10,300	1,755,250	6961	2246	1093	120p

eases with complex abnormalities is often unsatisfactory; see Tables 2 and 3. This is mainly due to the fact that reconstructing abnormalities from echocardiogram videos is more challenging, as it requires considering the local structural and motion information presented by the heart.

To this end, we present a novel approach called CardiacNet for the assessment of various cardiac diseases. Our key assumption is that once the model is equipped with the capability to accurately reconstruct abnormalities from normal cases, it can gain a better understanding of the diseases in terms of their local structural details and motion changes, and vice versa. To achieve it, our CardiacNet consists of three key components: **(1) Consistency Deformation Codebook (CDC)** is designed to simulate the reconstruction process between normal and abnormal cases, enabling the model to learn the local structural abnormalities and motion changes associated with the diseases. **(2) Consistency Deformation Discriminator (CDD)** aims to improve the quality of reconstructed videos and maintaining spatiotemporal consistency with the real videos in a discriminative manner. It prevents the degradation of reconstruction results by preserving the cardiac motion characteristics and introduces regional discrimination to maintain the local consistency of cardiac structural information. **(3)** We introduce a bidirectional reconstruction network to facilitate the learning of feature distributions for both normal and abnormal cases. This approach enhances the reconstruction process, enabling us to establish the respective distributions and explicitly optimize the distributions of these two distinct groups.

We evaluate our method in EF prediction using two publicly available datasets, CAMUS [13] and EchoNet [25], which are the only publicly available echocardiogram video datasets for cardiac disease assessment. To comprehensively evaluate the performance of CardiacNet across a diverse array of cardiac diseases, we introduce two benchmark datasets, namely **CardiacNet-PAH** and **CardiacNet-ASD**, specifically designed for PAH and ASD assessment. A detailed comparison between our datasets and the public datasets is provided in Table 1. Experimental results demonstrate that CardiacNet achieves state-of-the-art performance in three cardiac disease assessment tasks, including EF, PAH, and ASD.

To summarize, the main contributions of this paper are:

- We have constructed two benchmark datasets, the CardiacNet-PAH and the CardiacNet-ASD, specifically designed for cardiac disease assessment using echocardiogram videos.
- CardiacNet is a novel approach that can capture local structural details and cardiac motion changes, enabling accurate assessment of cardiac diseases.

- Our CardiacNet surpasses prior work in classifying PAH and ASD with an improvement of 2.1% and 5.0% in accuracy. The CardiacNet also achieves a relative reduction of 5.2% compared to prior arts in the EF prediction task.

2 Related Works

2.1 Diseases Analysis on Different Modalities

Currently, deep learning-based medical image representation learning on different modalities, such as CT, MRI and X-ray, typically use the reconstruction approach [11, 15–17, 33, 40]. They usually learn the distribution from the control normal group and detect out-of-distribution abnormalities with significant low-level details, such as tumours and bone fractures. These approaches struggle to differentiate between the complex abnormalities of a specific disease, as the model focuses more on reconstructing each sample independently but lacks consideration across data samples. Gradient-weighted Class Activation Mapping [23, 48, 50] can highlight the classification decision of feature maps from the network. Attention [39, 41] aims to highlight the out-of-distribution feature for abnormalities by introducing the attention mechanism. [23, 48, 50] use the anatomy-guided attention module to describe the confidence of the location of anomalies and treat them as explicit features to fine-tune the classification network. However, these methods rely on the classification backbone that is susceptible to noise and lacks the precision to accurately locate anomalous regions. The above methods serve for other medical modalities mainly focusing on medical images with significant lesions and pathology but lack consideration of both temporal and spatial information of cardiac data.

2.2 Cardiac Diseases Assessment from Echocardiogram Videos

For echocardiogram video, the anomaly analysis can be grouped into anomalies classification [14, 18] and anomalies visualization [18, 34], which offer baselines for adapting activation maps visualization of classification [18], and reconstruction-based [34] methods. [14] first adapt the regional myocardial wall motion tracking to detect abnormalities and quantify cardiac function. However, it only focuses on a single cardiac structure and ignores other information. [34] make the first attempt to reconstruct echocardiogram videos of normal groups from abnormal cases for congenital heart defect (CHD) detection. Yet this method barely considers prior knowledge of cardiac morphology. The lack of feature constraints also leads to the low quality of the reconstructed image. CAMUS [13] and EchoNet-Dynamic [13] are pioneer research that first proposes the echocardiogram video datasets for cardiac function evaluation. They also introduce the segmentation information for reference to predict the ejection fraction score. However, this task only reveals one of the cardiac functional parameters that is not able to classify the other cardiac diseases.

To overcome those problems, we thus propose a novel CardiacNet that builds a consistent relationship of morphological deformation between normal and abnormal cases by introducing prior knowledge of cardiac, which helps enable more

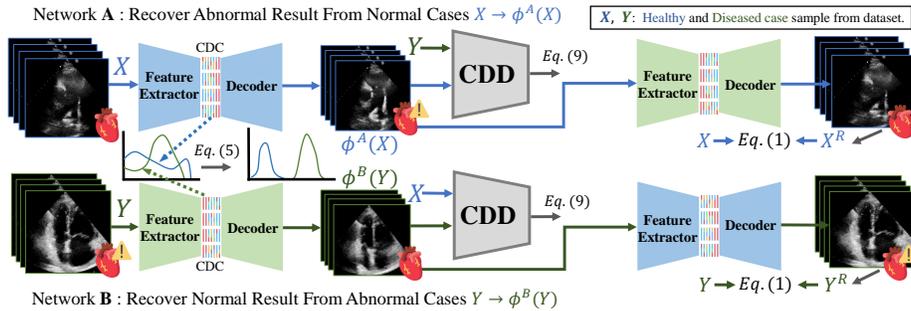


Fig. 2: The overview of our CardiacNet, sample normal case X and abnormal case Y , reconstruct the corresponding abnormal and normal results through networks $\phi^A(\cdot)$ and $\phi^B(\cdot)$, respectively. The Consistency Deformation Discriminator (CDD) is introduced to retain high reconstruction quality and allow the reconstruction results to be consistent with actual cases.

accurate evaluation in different tasks. Our new CardiacNet-PAH and CardiacNet-ASD datasets provide two different cardiac diseases related to cardiac morphology abnormalities and motion dysfunction.

3 Methodology

In this section, we introduce the CardiacNet as shown in Fig. 2. Hierarchically, CMT consists of the bidirectional reconstruction pipeline that simulates the deformation process from “normal” to “abnormal” cases and the reverse process. The Consistency Deformation Codebook (CDC) is designed to formulate deformation processes, allows the network to identify patterns of cardiac structures and motion from data samples with a specific cardiac disease, expect reconstructed results to match the corresponding features of real samples. The introduction of module Consistency Deformation Discriminator (CDD) is to discriminate whether reconstructed results are consistent with real data samples both spatially and temporally. It also guarantees high-quality echocardiogram video reconstruction.

3.1 Bidirectional Reconstruction Network

As shown in Fig. 2, two independent networks $\phi^A(\cdot)$ and $\phi^B(\cdot)$ with the same type of feature extractor, deformation codebook and decoder, respond to the reconstruction process of cases between “normal” and “abnormal”. Using the echocardiogram video input $X \in \mathbb{R}^{N \times H \times W \times 3}$ sample from the normal set as an example, where N is the total frame number of the X . First, divide each frame of X into regular non-overlapping patches and perform masking with the randomly sampled subset of patches. Then, compute reconstructed abnormal result $\phi^A(X)$. In the final, network $\phi^B(\cdot)$ transforms $\phi^A(X)$ as X^R to the normal result as the same as input X . With the $L1$ loss as our supervised reconstruction loss as follows:

$$\mathcal{L}_{\text{recon}}(X, X^R) = \|X - X^R\|_1, \quad (1)$$

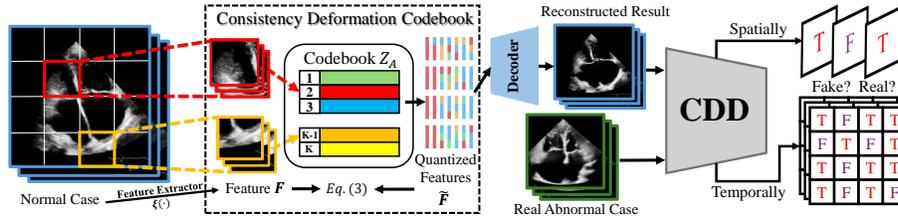


Fig. 3: The description of the one-way process of our CardiacNet. The encoded feature F of the normal case will be quantized by the deformation codebook Z as quantized feature \tilde{F} . The decoder then recovers \tilde{F} to the reconstructed abnormal result. Accompanied by the abnormal case sampled from the dataset, the Consistency Deformation Discriminator (CDD) is introduced to improve the consistency between reconstructed results and actual samples with regional discrimination.

where $\|\cdot\|_1$ indicate L_1 norm. The reconstructed abnormal $\phi^A(X)$ and real case sampled from the abnormal set will be discriminated by the CDD and compute the adversarial loss \mathcal{L}_{adv} . The process for reconstructing normal results from abnormal cases shares the same processing pipeline as the above description.

3.2 Consistency Deformation Codebook

Section 1 and research [3, 4, 24, 26, 37] illustrated that human cardiac remains structurally and morphologically similar, the lesions of cardiac diseases and its motion are generally dominated by specific locations of main structures and their substructures (refer to Fig. 1). With a large number of medical cases confirmed by experts, the pattern of cardiac structures and motion between normal and abnormal are able to be learned across samples. Hence, the main goal of the **Consistency Deformation Codebook (CDC)** is designed to formulate the pattern from medical cases. We hypothesise that the network understands the representation of a specific disease that can also reconstruct normal from abnormal or its reverse process. Hence, to simulate such behaviours, 1). The proposed CDC constructs the regional representation for different cardiac structures in order to maintain the temporal and spatial properties consistent between original and reconstructed echocardiogram videos. 2). To differentiate the deformation from “normal” to “abnormal” and its reversed process, we use the transport distance to distribute the discrepancy of two different distributions from large data samples and optimize the CDC module of network $\phi^A(\cdot)$ and $\phi^B(\cdot)$.

Consistent Deformation Encoding. As the pipeline described in Section 3.1 and Fig. 3, the CDC receives the feature map F encoded by the feature extractor $\xi(\cdot)$ of the network from the input. As discussed above, formulating the deformation process regionally is a more natural fit in echocardiogram videos. In order to perform this approach, we discretise the continuous feature F and reconstruct its latent representation regionally in a vector quantization manner. We first rewrite the F as $F = \{F_{n,i,j}\}_{n,i,j}^{N \times h \times w} \subset \mathbb{R}^d$ for querying the codebook entries $\mathcal{Z} = \{Z_k\}_{k=1}^K \subset \mathbb{R}^d$, where K is the total length of entries. In this step, directly applying the codebook [2] for quantizing videos disrupts temporal consistency.

Thus, we add learnable position encoding $\mathcal{P} = \{P_n\}_{n=1}^N \subset \mathbb{R}^d$ to feature maps F along the temporal dimension, which guarantees temporal consistency locally and globally. Given a subsequent element-wise quantization $\sigma(\cdot)$, we generate reconstructed abnormal feature $\sigma(F)$ as following:

$$\tilde{F} = \sigma(F, \mathcal{Z}, \mathcal{P}) := \left(\arg \min_{Z_k \in \mathcal{Z}} \left\| (F_{n,i,j} + P_n) - Z_k \right\|_2^2 \right)_{n,i,j} \in \mathbb{R}^{t \times h \times w \times d}. \quad (2)$$

For the loss of CDC, following the previous research [2, 44], we end-to-end train the CDC via Equation 3.

$$\mathcal{L}_q(\xi(I), \tilde{F}) = \left\| sg[\xi(I)] - \tilde{F} \right\|_2^2 + \lambda \cdot \left\| sg[\tilde{F}] - \xi(I) \right\|_2^2, \quad (3)$$

where I , $sg[\cdot]$, and λ denote the input of network $\phi(\cdot)$, the stop-gradient operation, and the factor of the second loss item that is set as 0.25. Equation 3 guarantees the network commits to the \mathcal{Z} since its dimensionless embedding space may grow arbitrarily during training. For the optimization of the CDC, we use the exponential moving average (EMA) method to update the codebook \mathcal{Z} as the following equation:

$$\mathcal{Z}'_{\text{new}} = (1 - \omega) \cdot \mathcal{Z} + \omega \cdot \mathcal{Z}_{\text{new}}, \quad (4)$$

where ω is the weight for updating the current codebook that is set as 0.01.

Optimal Transport Distance Optimization. The codebook of module CDC in Section 3.2 is proposed to formulate the pattern of the deformation process through all data samples from the dataset. To distinguish the distribution of normal and abnormal sets that are learned by codebooks of network $\phi^A(\cdot)$ and $\phi^B(\cdot)$, a more intuitive way is to use relative entropy to represent how one probability distribution differs from another. In this paper, we adopt the optimal transport measurement and expect to maximize the distance of deformations between the normal and abnormal sets. As shown in Fig. 4, networks $\phi^A(\cdot)$ and $\phi^B(\cdot)$ response for feature encoder, compute F_X , F_Y from normal case X and abnormal case Y . Implicitly, we can directly optimize the distance between codebooks of $\phi^A(\cdot)$ and $\phi^B(\cdot)$, which optimizes empirical distributions from entries instead of all data samples corresponding to each category from the dataset. However, due to the entries of each codebook being irrelevant and redundant, an entry in the same position of different codebooks is non-matching and non-equivalent, which can always be easily optimized to maximize their distance within a few iterations. Such orderless matching thus will invalidate the optimization, which indicates implicit optimization is not suitable for our approach.

To tackle this problem, we build two updated memory banks to store features encoded by CDC for normal and abnormal cases iteratively, which approximates the distribution of space of data samples. Similar to using EMA to update the codebook \mathcal{Z}^A and \mathcal{Z}^B , in memory banks \mathcal{M}^A , \mathcal{M}^B , we replace the ancestral features with current descendant features from Equation 2 and update the centroid with the EMA approach. Then, explicitly compute the transport distance

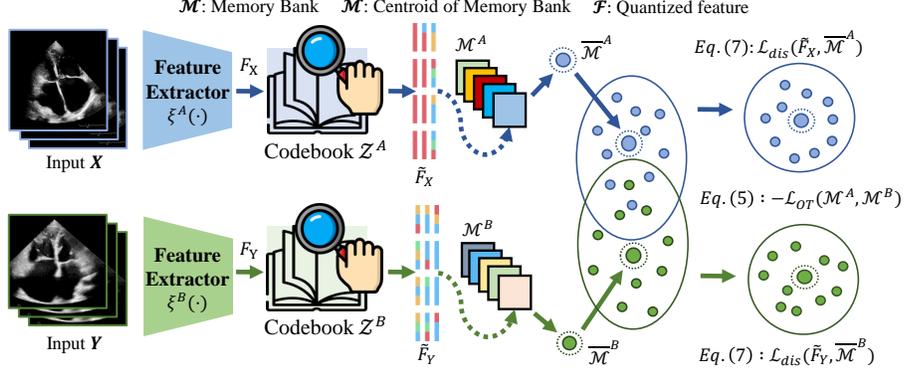


Fig. 4: The optimal transport distance optimization between two networks $\phi^A(\cdot)$ and $\phi^B(\cdot)$. Memory banks \mathcal{M}^A and \mathcal{M}^B store the features of normal and abnormal data samples, respectively. The loss $\mathcal{L}_{\text{OT}}(\mathcal{M}^A, \mathcal{M}^B)$ makes these two distributions keep away from each other. Losses $\mathcal{L}_{\text{dis}}(\tilde{F}_X, \bar{\mathcal{M}}^A)$, and $\mathcal{L}_{\text{dis}}(\tilde{F}_Y, \bar{\mathcal{M}}^B)$ make representations of clusters more consistent.

between codebooks by using Wasserstein distance [8] with Sinkhorn iteration [1], which formulate as the following equation:

$$\mathcal{L}_{\text{OT}}(\mathcal{M}^A, \mathcal{M}^B) = \sum_{i=1}^d \sum_{j=1}^J \left\| \mathcal{M}_{j,i}^A - \mathcal{M}_{\pi^i(j),i}^B \right\|_2^2, \quad (5)$$

where J denotes the number of samples stored in the memory bank \mathcal{M} , $\mathcal{M}_{j,i}$ denote the i -th dimension of j -th sample in \mathcal{M} . The $\pi^i(\cdot)$ is a mapping function to minimize the transport distance of samples between two memory banks as the following:

$$\pi^i = \underset{\pi}{\operatorname{argmin}} \sum_{j=1}^J \left\| \mathcal{M}_{j,i}^A - \mathcal{M}_{\pi(j),i}^B \right\|_2^2. \quad (6)$$

Additionally, we minimize the distance between the current quantized feature and the centroid of the corresponding memory bank. Hence, we use the representative centroid that averages the features of all samples in \mathcal{M} as $\bar{\mathcal{M}}$. The centroids $\bar{\mathcal{M}}$ then is used for measuring the discrepancy with the quantized feature \tilde{F} (defined in Equation 2), the loss is formulated in the following form:

$$\mathcal{L}_{\text{dis}}(\tilde{F}, \bar{\mathcal{M}}) = \left\| \tilde{F} - \bar{\mathcal{M}} \right\|_2^2. \quad (7)$$

Similarly, the same operation will be conducted for the abnormal input Y . The overall loss \mathcal{L}_{CDC} for optimizing the CDC is:

$$\begin{aligned} \mathcal{L}_{\text{CDC}} = & \mathcal{L}_q(F_X^A, \tilde{F}_X^A) + \mathcal{L}_q(F_Y^B, \tilde{F}_Y^B) \\ & + \mathcal{L}_{\text{dis}}(\tilde{F}_X^A, \bar{\mathcal{M}}^A) + \mathcal{L}_{\text{dis}}(\tilde{F}_Y^B, \bar{\mathcal{M}}^B) + \mathcal{L}_{\text{OT}}(\mathcal{M}^A, \mathcal{M}^B), \end{aligned} \quad (8)$$

where $F_I^\theta = \xi^\theta(I)$ and $\tilde{F}_I^\theta = \sigma(F_I^\theta, \mathcal{Z}^\theta, \mathcal{P}^\theta)$ for $\theta \in \{A, B\}$, and $I \in \{X, Y\}$ for paired (X, Y) input.

3.3 Consistency Deformation Discriminator

The introduction of CDD ensures the reconstructed echocardiogram videos remain consistent in their spatially and temporally visual properties, such as textures and colors. Also, the discriminator acts as an adversary that forces the reconstructed results to conform with the real data in semantic properties, such as structural abnormalities and motion dysfunction of a specific cardiac disease. Hence, the CDD consists of two discriminators, denoted as $\eta^S(\cdot)$ and $\eta^T(\cdot)$, which discriminate reconstructed results and real samples. The $\eta^T(\cdot)$ for spatial consistency discriminates every single frame of a video while the $\eta^S(\cdot)$ responds to the temporal consistency that takes the whole video as input. Using the reconstruction process from normal to abnormal as an example, we let $\{\hat{X}_n\}_{n=1}^N = \phi^A(X)$ and $\{\hat{Y}_n\}_{n=1}^N = \hat{Y}$ to represent the reconstructed video and real abnormal video, respectively. As shown in Fig. 3, globally, we use the $\eta^T(\cdot)$ to discriminate the whole reconstructed video $\phi^A(X)$ and sampled real video \hat{Y} , as the first term in Equation 9. The $\eta^T(\cdot)$ takes each frame of $\phi^A(X)$ and \hat{Y} in order as an image pair for the spatial discrimination as the second term in Equation 9.

Locally, we need to guarantee that each region of cardiac can also conduct high-quality reconstruction as well as remain consistent with real cases. For example, for the process of reconstructing normal X to abnormal $\phi^A(X)$, the discrepancy of motion between reconstructed results $\phi^A(X)$ and real abnormal sample Y should remain consistent for a person. Hence, we first convert \hat{Y} and $\phi^A(X)$ to non-overlap patches as $\{\hat{Y}_{i,j}\}_{i=1,j=1}^{h,w}, \{\hat{X}_{i,j}\}_{i=1,j=1}^{h,w} \in \mathbb{R}^{N \times \frac{H}{h} \times \frac{W}{w} \times 3}$, where $\frac{H}{h}, \frac{W}{w} \in \mathbb{Z}^+$, W and H is the width and height of input images, w and h is the size of width and height of the feature map. The overall adversarial loss for global and local discrimination can be formulated as the following Equation:

$$\begin{aligned} \mathcal{L}_{\text{adv}}(\phi^A(X), Y) &= (\log(\eta^T(\phi^A(X))) + \log(1 - \eta^T(Y))) \\ &+ \sum_{n=1}^t \left[\log(1 - \eta^S(\hat{X}_n)) + \log(\eta^S(\hat{Y}_n)) \right] \\ &+ \sum_{i=1,j=1}^{h,w} \left[\log(1 - \eta^T(\hat{X}_{i,j})) + \log(\eta^T(\hat{Y}_{i,j})) \right]. \end{aligned} \quad (9)$$

To address the necessity of both global and local discrimination, we conduct the ablation study as shown in Section 4.4 and Table 4. For the overall adversarial loss, according to Equation 9, we have $\mathcal{L}_{\text{CDD}} = \mathcal{L}_{\text{adv}}(\phi^A(X), Y) + \mathcal{L}_{\text{adv}}(\phi^B(Y), X)$. In the final, applying the End-to-End training for the CMT and combining the loss from CDC and CDD, the overall loss of our CardiacNet is $\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{CDC}} + \mathcal{L}_{\text{CDD}} + \mathcal{L}_{\text{recon}}(X, X^R) + \mathcal{L}_{\text{recon}}(Y, Y^R)$.

4 Experiments

4.1 Dataset

We evaluate our method on three datasets, including two public datasets CAMUS [13] and Echonet-Dynamic [25], as well as our collected dataset CardiacNet-PAH and CardiacNet-ASD.

CardiacNet-PAH and CardiacNet-ASD. We collect datasets from four collaborating hospitals. To guarantee all echocardiogram videos are standards-compliant, each case underwent a video from the apical four-chamber heart (A4C) view are collected, annotated and approved by 5-6 experienced physicians. Ethically, we strictly adhere to the ethical standards of medical research and ensure that the local ethics committee approves all image data collection and experiments. As shown in Table 1, the CardiacNet-PAH consists of 496 cases for classifying Pulmonary Arterial Hypertension (PAH), and the diagnosis of patients is accessed and approved through Right Heart Catheterization measurement. In CardiacNet-ASD, 231 cases for classifying the Atrial Septal Defect (ASD) are diagnosed and annotated by experienced physicians. The resolution of each video is either 800×600 or 1024×768 , depending on the type of scanner (Philips or HITACHI). A total of 727 videos are collected, and each video consists of over 100 frames, covering at least two heartbeat cycles. We also collect Pixel-level annotations of cardiac structure for reconstruction evaluation, including masks for the left ventricle (LV), right ventricle (RV), left atrium (LA), and right atrium (RA) in the A4C view. Five frames are provided with pixel-level annotation masks for each video.

CAMUS [13] and EchoNet-Dynamic [25]. CAMUS consists of 500 echocardiogram videos with pixel-level annotations for the left ventricle, myocardium, and left atrium. EchoNet-Dynamic [25] (EchoNet) is the largest echocardiogram video dataset, including 10,030 videos. Both datasets annotated 2 frames (end diastole and end systole) of left ventricle segmentation. The Ejection Fraction (EF) score is provided for each video for the regression task. In this paper, we follow the [21] that use cases in CMAUS and EchoNet with $EF \leq 50\%$ as the abnormal group while $EF \geq 55\%$ as the normal group for classification class.

4.2 Implementation Details

Training. The backbone of our methods is built on the generative network [2]. We trained the model using the Adam optimizer with a weight decay of $1e^{-3}$ and a momentum of 0.9. The model was trained for a total of 1,000 epochs with an initial learning rate of $2.25e^{-4}$, and the learning rate was decreased by a factor of 0.1 for every 400 epochs. The batch size was set to 2 in our experiment. For spatial data augmentation, each frame was resized to 144×144 and then randomly cropped to 112×112 . The frames were also randomly flipped vertically and horizontally. For temporal data augmentation, we randomly selected 48 continuous frames from an echocardiogram video and sampled 16 frames as input equidistantly. The CardiacNet was split in a ratio of 8:1:1 for training, validation and testing. For the CAMUS and EchoNet datasets, we follow the same data argumentation recipe as our CardiacNet. We also follow the default dataset split provided by the official setting [13] and [25].

Inference and Testing. During this stage, we took the feature extractor $\xi^A(\cdot)$ of network $\phi^A(\cdot)$ saved from the final iteration as our testing model. For car-

diac disease assessment tasks, classification and regression, we first freeze the parameter of the feature extractor in the trained model. Then, for each input, we flattened the feature and fine-tuned different tasks with a single Linear layer. We report the final results and perform visualization on the testing set. During this stage, we do not conduct any argumentation for input echocardiogram video except resize frames to 144×144 and apply center cropping to 112×112 . For the length of input videos in inference, The number of input frames is 16, and the sample rate is 4. For evaluating the reconstruction result, we trained a segmentation network according to the segmentation annotation presented by the dataset. During inference, the reconstruction result from pre-trained network $\phi^A(\cdot)$ will be input to the segmentation network and perform the evaluation.

Evaluation Metrics. For PAH, ASD and EF classification, we use the Area Under the ROC Curve (AUC) and classification accuracy (ACC) to evaluate the performance of trained networks in classifying anomalies. We predict the EF values and report the Mean Absolute Error (MAE) for CAMUS and Echonet datasets that evaluate the Ejection Fraction (EF) score. In order to evaluate the reconstruction quality, we use Fréchet Inception Distance (FID) to evaluate the quality of recovery images. For ASD, we also introduce the DICE score to evaluate whether recovered images are consistent with the original image in the ventricles and atrium of cardiac structures. This is due to the recovery from ASD to normal does not affect the volume of cardiac structures. For each method, we also compare their efficiency by reporting the inference time, the number of parameters (MParams) and Tera-Flops (TFlops).

4.3 Results

Result on CardiacNet-PAH and CardiacNet-ASD. Table 2 illustrates the comparison result of PAH classification results in CardiacNet-PAH. We currently categorise open-source methods into classification/regression models and reconstruction-based models. The AUC-ROC and ACC illustrate the performance of models in distinguishing normal and abnormal cases. Our CardiacNet achieves 89.32% and 85.71% in AUC-ROC and ACC, respectively, while the HiFuse [9] reaches the second-best results with 84.11% and 83.67%, where CardiacNet surpass by +5.21% and +2.04%. Indicating our method can outperform other methods by a considerable margin. For reconstructed image quality evaluation, compared to the Wolleb et al. [45] reaches 16.12 in FID score, our method can achieve 14.73, which shows that our method can perform better reconstruction quality in echocardiogram videos.

Compared to PAH classification, classifying ASD is an easier task due to ASD presenting more significant morphological anomalies. For the classification performance in CardiacNet-ASD, the AUC-ROC and ACC presented by our method are 91.24% and 89.63%, outperforming the best baseline DeepGuide [20] by a margin of +6.22% and +4.84%. As illustrated in Table 2, CardiacNet achieves 15.22 in FID score with 0.56 improvement than the second best method Wolleb

[†]One step TFlops of denoising. A total of 1000 steps are used in inference.

Table 2: The result of classification in our CardiacNet-PAH and CardiacNet-ASD, reporting results in metrics FID, AUC-ROC (%) and ACC (%). For ASD, with DICE (%) score to evaluate the segmentation accuracy of reconstructed images compared with ground truth. The classification networks do not reconstruct the image, and the FID is not provided for these approaches. Underline denotes the second-best result.

Methods	Datasets						Efficiency			
	CardiacNet-PAH			CardiacNet-ASD			Time↓	MParams↓	TFlops↓	
	FID↓	AUC-ROC↑	ACC↑	FID↓	DICE↑	AUC-ROC↑				ACC↑
<i>Classification Network</i>										
ResNet3D [7]	-	77.32	71.43	-	-	72.25	75.86	2.479	47.02	0.202
AGXNet [48]	-	76.09	72.41	-	-	76.52	72.41	2.873	12.31	0.210
EchoNet [25]	-	81.63	80.95	-	-	83.62	82.75	2.653	33.19	0.848
DeepGuide [20]	-	82.45	81.63	-	-	<u>85.02</u>	<u>84.79</u>	3.780	15.60	0.748
DiffMIC [40]	-	81.73	79.59	-	-	82.81	81.48	1182	88.56	38.58 [†]
HiFuse [9]	-	<u>84.11</u>	<u>83.67</u>	-	-	81.08	79.31	3.183	135.7	5.106
<i>Reconstruction-Based Methods</i>										
Vanilla GAN [2]	18.90	52.37	46.15	19.07	63.55	60.54	58.62	2.221	12.95	0.842
DAE [11]	16.39	58.91	57.69	15.38	65.80	54.09	53.77	1534	159.4	78.08 [†]
VTGAN [10]	17.66	58.32	51.72	18.10	65.13	70.92	68.97	38.50	243.3	1.423
Att. UNet [39]	18.42	57.29	55.17	18.95	64.30	69.81	62.06	2.621	34.88	4.081
Wolleb et al. [45]	<u>16.12</u>	70.42	67.35	<u>15.78</u>	68.61	67.88	65.51	1488	89.87	45.13 [†]
DeScarGAN [46]	16.59	64.21	71.42	17.04	68.52	71.33	68.97	2.756	8.528	2.756
Diff-SCM [35]	15.57	64.23	61.22	16.37	63.26	69.23	70.83	1295	53.41	40.37 [†]
CyTran [33]	16.40	72.69	69.38	16.93	70.21	<u>74.35</u>	72.41	2.769	1.191	0.125
CardiacNet (Ours)	14.73	89.32	85.71	15.22	73.52	91.24	89.63	4.523	28.34	7.949

Table 3: The result of ejection fraction regression and abnormal classification in publicly CAMUS [13] and EchoNet [25] dataset. Reporting results in metrics FID, AUC-ROC (%) and ACC (%), with Mean Absolute Error (MAE) of ejection fraction score regression for CAMUS and EchoNet. The classification/regression networks do not reconstruct the image, and the FID is not provided for these approaches. Underline denotes the second-best result.

Methods	Datasets							
	CAMUS			EchoNet				
	FID↓	MAE↓	AUC↑	ACC↑	FID↓	MAE↓	AUC↑	ACC↑
<i>Classification / Regression Network</i>								
ResNet3D [7]	-	7.59	70.34	68.00	-	5.44	78.80	75.44
AGXNet [48]	-	6.91	76.58	72.00	-	5.17	78.46	80.02
DeepGuide [20]	-	6.72	79.66	74.00	-	4.70	84.33	79.59
EchoNet [25]	-	<u>6.30</u>	<u>80.75</u>	<u>76.00</u>	-	4.22	83.19	81.52
HiFuse [9]	-	6.34	80.26	76.00	-	<u>4.08</u>	<u>85.73</u>	<u>82.41</u>
<i>Reconstruction-Based Methods</i>								
Vanilla GAN [2]	17.24	12.59	65.11	66.00	17.36	20.23	50.18	50.60
VTGAN [10]	16.95	13.72	61.62	56.00	15.83	12.87	61.56	61.05
Att. UNet [39]	17.72	9.48	65.60	62.00	16.44	8.25	65.09	61.92
CyTran [33]	15.82	8.52	66.42	66.00	15.07	7.59	68.45	66.53
DeScarGAN [46]	15.56	6.80	73.24	68.00	14.19	7.23	73.24	71.08
Wolleb et al. [45]	15.17	8.06	75.96	74.00	13.18	8.50	72.38	69.57
CardiacNet (Ours)	14.64	5.97	83.09	80.00	<u>13.25</u>	3.83	86.52	84.70

et al. [45]. Also, to evaluate that the reconstructed image is consistent in volume sizes of different cardiac structures, our method achieves the best Dice score of 73.52%, while other methods are significantly below 70%.

Result on CAMUS and EchoNet. As shown in Table 3 in columns CAMUS and EchoNet, for the regression task of EF score prediction in both datasets, results achieved by our method are considerably better than others, with 5.97 and 3.83 MAE in the regression task. In contrast, the second best method, HiFuse [9], reaches only 6.34 and 4.08 in MAE, respectively. Illustrates our method CardiacNet is able to learn the better representation for the regression task. For disease

Table 4: Effectiveness of CDC and CDD. Results report in CardiacNet-PAH.

CDC	CDD	Results		
		FID	AUC	ACC
X	X	18.90	52.37	46.15
✓	X	16.82	80.27	79.59
X	✓	17.09	52.46	53.84
✓	✓	14.73	89.23	85.71

Table 5: Ablation study of Positional Encoding and Optimal Transport in **only CDC module**.

Pos. Encode	Opt. Trans	Results		
		FID	AUC	ACC
X	X	18.90	52.37	46.15
✓	X	17.41	62.44	65.38
X	✓	18.06	78.39	75.51
✓	✓	16.82	80.27	79.59

Table 6: Ablation study of Global and Local discriminator in CDD module (Enabling CDC).

Global. CDD	Local. CDD	Results		
		FID	AUC	ACC
X	X	16.82	80.27	79.59
✓	X	15.62	82.41	83.67
X	✓	15.41	84.57	81.63
✓	✓	14.73	89.23	85.71

classification, the AUC-ROC and ACC of our method in CMUAS are 83.09% and 79.11%, respectively, while reaching 86.52% and 84.70% in EchoNet. The second best method is HiFuse with the AUC-ROC and ACC of 80.26% and 76.13% in CAMUS as well as 85.73% and 82.41% in EchoNet, respectively. Results illustrate our method is more accurate in classifying patients with abnormal left ventricular endocardium in both end-diastole (ED) and end-systole (ES). Compared with other methods in reconstructing high-quality videos, our method can achieve the FID score of 14.64 and 13.25 while Wolleb et al. [45] achieve 15.17 and 13.18 in CMUAS and EchoNet datasets, with the higher reconstruction quality in EchoNet dataset.

4.4 Ablation Study

Consistency Deformation Codebook. As shown in Table 5, the ablation study of the CDC module consists of positional encoding and optimal transport. Compared to the results of disabling these two modules, the position encoding for temporal consistency can enhance reconstruction quality by around 1.34 in FID and classification accuracy by around 20%. The improvement of the above two numbers contributed by optimal transport is around 0.84 and 30%. These results show both the positional embedding and optimal transport are efficient and can help CardiacNet to learn the better representation of cardiac diseases. Furthermore, we visualize the embedding features of PAH patients and normal cases in Fig 5. Our reconstruction network produces embedding features without using additional layers, which shows that our CDC can help distinguish cardiac structural and motion abnormalities.

Consistency Deformation Discriminator. As shown in Table 6, both global and local discriminators can contribute to the CDD module. Due to their constraint of the spatial and temporal consistency within patches (see section 3.3), the CDD brings improvement in reconstructed image quality. Using only the global or local discriminator leads to significant degradation in the FID score and classification accuracy. The ablation study in Table 4 shows the combination of CDD, CDC, and CardiacNet achieves the best performance in both reconstruction and classification.

The Visualization of Reconstruction Cases. As shown in Fig. 6, our method is able to reconstruct the possible “normal” images from abnormal cases compared with other reconstruction methods. Our reconstructed images remain high

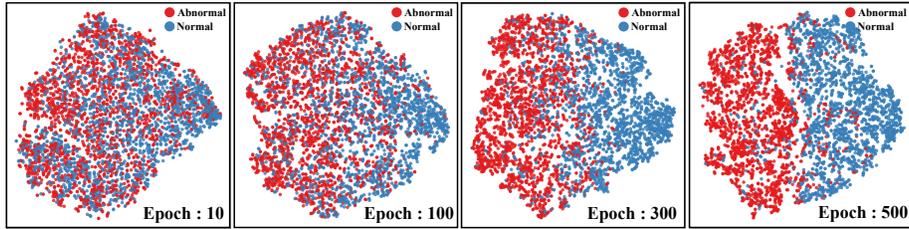


Fig. 5: The visualization of t-SNE results between learned embedding of normal and abnormal cases by our CardiacNet in epochs 10, 100, 300 and 500.

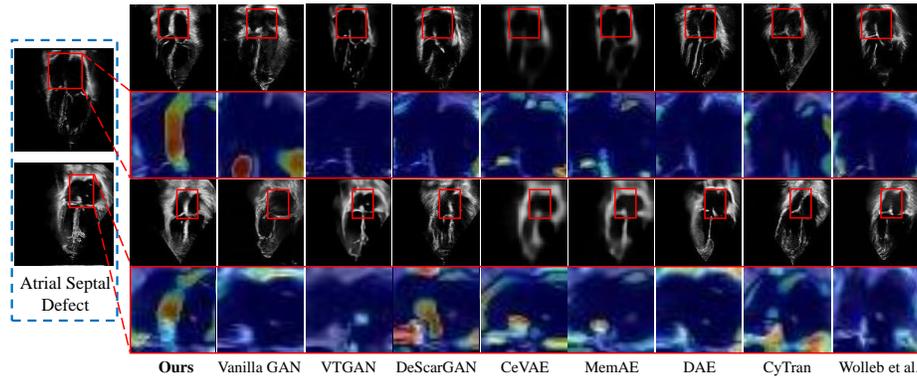


Fig. 6: The visualization case of recovery results across ours and eight different reconstruction-based methods [6, 10, 11, 33, 38, 39, 45, 46, 52]. We use cases from patients with **Atrial Septal Defect** (ASD). We let experienced physicians annotate possible abnormal areas and visualize the difference by using the heatmap. (Best view in colour)

quality and can provide more reasonable visualization results that are approved by experienced physicians. As shown in two different cases, the disappearance of the atrial septum and the abnormal right atrial volume can be distinguished and recovered while maintaining the reconstruction quality.

5 Conclusion

In this paper, we first proposed a novel CardiacNet for learning the morphological abnormalities and motion dysfunction of cardiac disease through echocardiogram videos. We introduce a new benchmark dataset that includes two different types of cardiac diseases as well as cardiac structure segmentation. All cases are annotated and confirmed by experienced physicians, which can significantly contribute to the medical image analysing community and further the development in detecting morphological abnormalities and motion dysfunction for cardiac diseases. In our future study, we will further our exploration in more fine-grained echocardiogram video reconstruction that enables symptom grading for diseases with the visualization of morphological lesions. Moreover, we will make attempts to involve other state-of-the-art techniques, such as Large language models (LLMs) and multi-modality fusion, to generate more precise and robust results.

Acknowledgement

This work was supported by a research grant from the Beijing Institute of Collaborative Innovation (BICI) under collaboration with the Hong Kong University of Science and Technology under Grant HCIC-004 and a project of Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone (HZQB-KCZYB-2020083).

References

1. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. *NIPS* **26** (2013) [8](#)
2. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: *CVPR*. pp. 12873–12883 (2021) [6](#), [7](#), [10](#), [12](#)
3. Ganame, J., Mertens, L., Eidem, B.W., Claus, P., D’hooge, J., Havemann, L.M., McMahon, C.J., Elayda, M.A.A., Vaughn, W.K., Towbin, J.A., et al.: Regional myocardial deformation in children with hypertrophic cardiomyopathy: morphological and clinical correlations. *European heart journal* **28**(23), 2886–2894 (2007) [6](#)
4. Geske, J.B., Bos, J.M., Gersh, B.J., Ommen, S.R., Eidem, B.W., Ackerman, M.J.: Deformation patterns in genotyped patients with hypertrophic cardiomyopathy. *European Heart Journal–Cardiovascular Imaging* **15**(4), 456–465 (2014) [6](#)
5. Ghorbani, A., Ouyang, D., Abid, A., He, B., Chen, J.H., Harrington, R.A., Liang, D.H., Ashley, E.A., Zou, J.Y.: Deep learning interpretation of echocardiograms. *NPJ digital medicine* **3**(1), 10 (2020) [1](#)
6. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: *ICCV* (2019) [14](#)
7. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 6546–6555 (2018) [12](#)
8. Hörmander, F., Totaro, N., Waldschmidt, A.V.M.: *Grundlehren der mathematischen wissenschaften* 332, vol. 5. Springer (2006) [8](#)
9. Huo, X., Sun, G., Tian, S., Wang, Y., Yu, L., Long, J., Zhang, W., Li, A.: Hifuse: Hierarchical multi-scale feature fusion network for medical image classification. *Biomedical Signal Processing and Control* **87**, 105534 (2024) [2](#), [11](#), [12](#)
10. Kamran, S.A., Hossain, K.F., Tavakkoli, A., Zuckerbrod, S.L., Baker, S.A.: Vtgan: Semi-supervised retinal image synthesis and disease prediction using vision transformers. In: *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. pp. 3228–3238 (2021). <https://doi.org/10.1109/ICCVW54120.2021.00362> [2](#), [12](#), [14](#)
11. Kascenas, A., Pugeault, N., O’Neil, A.Q.: Denoising autoencoders for unsupervised anomaly detection in brain mri. In: *International Conference on Medical Imaging with Deep Learning*. pp. 653–664. PMLR (2022) [2](#), [4](#), [12](#), [14](#)
12. Lai, W.W., Mertens, L.L., Cohen, M.S., Geva, T.: *Echocardiography in pediatric and congenital heart disease: from fetus to adult*. John Wiley & Sons (2015) [1](#)
13. 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) [3](#), [4](#), [9](#), [10](#), [12](#)

14. Lin, X., Yang, F., Chen, Y., Chen, X., Wang, W., Chen, X., Wang, Q., Zhang, L., Guo, H., Liu, B., et al.: Echocardiography-based ai detection of regional wall motion abnormalities and quantification of cardiac function in myocardial infarction. *Frontiers in Cardiovascular Medicine* **9**, 903660 (2022) [4](#)
15. Lin, Y., Luo, Z., Zhao, W., Li, X.: Learning deep intensity field for extremely sparse-view cbct reconstruction. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. pp. 13–23. Springer Nature Switzerland (2023) [4](#)
16. Lin, Y., Wang, H., Chen, J., Li, X.: Learning 3d gaussians for extremely sparse-view cone-beam ct reconstruction (2024), <https://arxiv.org/abs/2407.01090> [4](#)
17. Lin, Y., Yang, J., Wang, H., Ding, X., Zhao, W., Li, X.: C²rv: Cross-regional and cross-view learning for sparse-view cbct reconstruction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 11205–11214 (June 2024) [4](#)
18. Liu, B., Chang, H., Yang, D., Yang, F., Wang, Q., Deng, Y., Li, L., Lv, W., Zhang, B., Yu, L., et al.: A deep learning framework assisted echocardiography with diagnosis, lesion localization, phenogrouping heterogeneous disease, and anomaly detection. *Scientific Reports* **13**(1), 3 (2023) [1](#), [4](#)
19. Lu, Y., Li, K., Pu, B., Tan, Y., Zhu, N.: A yolox-based deep instance segmentation neural network for cardiac anatomical structures in fetal ultrasound images. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2022) [1](#)
20. Mallya, M., Hamarneh, G.: Deep multimodal guidance for medical image classification. In: *MICCAI*. Springer (2022) [2](#), [11](#), [12](#)
21. McDonagh, T.A., Metra, M., Adamo, M., Gardner, R.S., Baumbach, A., Böhm, M., Burri, H., Butler, J., Čelutkienė, J., Chioncel, O., et al.: 2021 esc guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the european society of cardiology (esc) with the special contribution of the heart failure association (hfa) of the esc. *European heart journal* **42**(36), 3599–3726 (2021) [10](#)
22. Mcleod, G., Shum, K., Gupta, T., Chakravorty, S., Kachur, S., Bienvenu, L., White, M., Shah, S.B.: Echocardiography in congenital heart disease. *Progress in cardiovascular diseases* **61**(5-6), 468–475 (2018) [1](#)
23. Meena, T., Kabiraj, A., Reddy, P.B., Roy, S.: Weakly supervised confidence aware probabilistic cam multi-thorax anomaly localization network. In: *2023 IEEE 24th International Conference on Information Reuse and Integration for Data Science (IRI)*. pp. 309–314. IEEE (2023) [4](#)
24. Niemann, M., Liu, D., Hu, K., Cikes, M., Beer, M., Herrmann, S., Gaudron, P.D., Hillenbrand, H., Voelker, W., Ertl, G., et al.: Echocardiographic quantification of regional deformation helps to distinguish isolated left ventricular non-compaction from dilated cardiomyopathy. *European journal of heart failure* **14**(2), 155–161 (2012) [6](#)
25. 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) [1](#), [2](#), [3](#), [9](#), [10](#), [12](#)
26. Oxborough, D., Sharma, S., Shave, R., Whyte, G., Birch, K., Artis, N., Batterham, A.M., George, K.: The right ventricle of the endurance athlete: the relationship between morphology and deformation. *Journal of the American Society of Echocardiography* **25**(3), 263–271 (2012) [6](#)

27. Popp, R.L.: Echocardiographic assessment of cardiac disease. *Circulation* **54**(4), 538–552 (1976) [1](#)
28. Pu, B., Li, K., Chen, J., Lu, Y., Zeng, Q., Yang, J., Li, S.: Hfscdd: a hybrid neural network for fetal standard cardiac cycle detection in ultrasound videos. *IEEE Journal of Biomedical and Health Informatics* (2024) [1](#)
29. Pu, B., Lu, Y., Chen, J., Li, S., Zhu, N., Wei, W., Li, K.: Mobileunet-fpn: A semantic segmentation model for fetal ultrasound four-chamber segmentation in edge computing environments. *IEEE Journal of Biomedical and Health Informatics* **26**(11), 5540–5550 (2022) [1](#)
30. Pu, B., Lv, X., Yang, J., Guannan, H., Dong, X., Lin, Y., Shengli, L., Ying, T., Fei, L., Chen, M., et al.: Unsupervised domain adaptation for anatomical structure detection in ultrasound images. In: *Forty-first International Conference on Machine Learning* [1](#)
31. Pu, B., Wang, L., Yang, J., He, G., Dong, X., Li, S., Tan, Y., Chen, M., Jin, Z., Li, K., et al.: M3-uda: A new benchmark for unsupervised domain adaptive fetal cardiac structure detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 11621–11630 (2024) [1](#)
32. Pu, B., Zhu, N., Li, K., Li, S.: Fetal cardiac cycle detection in multi-resource echocardiograms using hybrid classification framework. *Future Generation Computer Systems* **115**, 825–836 (2021) [1](#)
33. Ristea, N.C., Miron, A.I., Savencu, O., Georgescu, M.I., Verga, N., Khan, F.S., Ionescu, R.T.: Cytran: Cycle-consistent transformers for non-contrast to contrast ct translation. *Neurocomputing* (2023). <https://doi.org/10.1016/j.neucom.2023.03.072> [2](#), [4](#), [12](#), [14](#)
34. Ryser, A., Manduchi, L., Laumer, F., Michel, H., Wellmann, S., Vogt, J.E.: Anomaly detection in echocardiograms with dynamic variational trajectory models. In: *Machine Learning for Healthcare Conference*. pp. 425–458. PMLR (2022) [1](#), [2](#), [4](#)
35. Sanchez, P., Kascenas, A., Liu, X., O’Neil, A.Q., Tsaftaris, S.A.: What is healthy? generative counterfactual diffusion for lesion localization. In: *MICCAI Workshop on Deep Generative Models*. pp. 34–44. Springer (2022) [12](#)
36. Sanjeevi, G., Gopalakrishnan, U., Pathinarupothi, R.K., Madathil, T.: Automatic diagnostic tool for detection of regional wall motion abnormality from echocardiogram. *Journal of medical systems* **47**(1), 13 (2023) [1](#)
37. Schäfer, M., Mitchell, M.B., Frank, B.S., Barker, A.J., Stone, M.L., Jagers, J., von Alvensleben, J.C., Hunter, K.S., Friesen, R.M., Ivy, D.D., et al.: Myocardial strain-curve deformation patterns after fontan operation. *Scientific Reports* **13**(1), 11912 (2023) [6](#)
38. Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis* **54**, 30–44 (2019) [14](#)
39. Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D.: Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis* **53**, 197–207 (2019) [4](#), [12](#), [14](#)
40. Silva-Rodríguez, J., Naranjo, V., Dolz, J.: Constrained unsupervised anomaly segmentation. *Medical Image Analysis* **80**, 102526 (2022) [2](#), [4](#)
41. Sun, D., Hu, Y., Li, Y., Yu, X., Chen, X., Shen, P., Tang, X., Wang, Y., Lai, C., Kang, B., et al.: Chamber attention network (can): Towards interpretable diagnosis of pulmonary artery hypertension using echocardiography. *Journal of Advanced Research* (2023) [1](#), [4](#)

42. Tseng, C.H., Chien, S.J., Wang, P.S., Lee, S.J., Pu, B., Zeng, X.J.: Real-time automatic m-mode echocardiography measurement with panel attention. *IEEE Journal of Biomedical and Health Informatics* (2024) [1](#)
43. Upton, M., Gibson, D., Brown, D.: Echocardiographic assessment of abnormal left ventricular relaxation in man. *Heart* **38**(10), 1001–1009 (1976) [1](#)
44. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. *NIPS* **30** (2017) [7](#)
45. Wolleb, J., Bieder, F., Sandkühler, R., Cattin, P.C.: Diffusion models for medical anomaly detection. In: *MICCAI*. pp. 35–45. Springer (2022) [11](#), [12](#), [13](#), [14](#)
46. Wolleb, J., Sandkühler, R., Cattin, P.C.: Descargan: Disease-specific anomaly detection with weak supervision. In: *MICCAI*. pp. 14–24. Springer (2020) [12](#), [14](#)
47. Yang, J., Ding, X., Zheng, Z., Xu, X., Li, X.: Graphecho: Graph-driven unsupervised domain adaptation for echocardiogram video segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 11878–11887 (2023) [1](#)
48. Yu, K., Ghosh, S., Liu, Z., Deible, C., Batmanghelich, K.: Anatomy-guided weakly-supervised abnormality localization in chest x-rays. In: *MICCAI*. pp. 658–668. Springer (2022) [2](#), [4](#), [12](#)
49. Zaman, F., Ponnappureddy, R., Wang, Y.G., Chang, A., Cadaret, L.M., Abdelhamid, A., Roy, S.D., Makan, M., Zhou, R., Jayanna, M.B., et al.: Spatio-temporal hybrid neural networks reduce erroneous human “judgement calls” in the diagnosis of takotsubo syndrome. *EClinicalMedicine* **40** (2021) [1](#)
50. Zhang, J., Xie, Y., Pang, G., Liao, Z., Verjans, J., Li, W., Sun, Z., He, J., Li, Y., Shen, C., et al.: Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection. *IEEE transactions on medical imaging* **40**(3), 879–890 (2020) [4](#)
51. Zheng, Z., Yang, J., Ding, X., Xu, X., Li, X.: Gl-fusion: Global-local fusion network for multi-view echocardiogram video segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 78–88. Springer (2023) [1](#)
52. Zimmerer, D., Kohl, S.A., Petersen, J., Isensee, F., Maier-Hein, K.H.: Context-encoding variational autoencoder for unsupervised anomaly detection. *arXiv preprint arXiv:1812.05941* (2018) [14](#)