

Feedback Attention to Enhance Unsupervised Deep Learning Image Registration in 3D Echocardiography

Md. Kamrul Hasan, Yihao Luo, Guang Yang[†], *Senior Member, IEEE*, Choon Hwai Yap[†]

Abstract—Cardiac motion estimation is important for assessing the contractile health of the heart, and performing this in 3D can provide advantages due to the complex 3D geometry and motions of the heart. Deep learning image registration (DLIR) is a robust way to achieve cardiac motion estimation in echocardiography, providing speed and precision benefits, but DLIR in 3D echo remains challenging. Successful unsupervised 2D DLIR strategies are often not effective in 3D, and there have been few 3D echo DLIR implementations. Here, we propose a new spatial feedback attention module (FBA) to enhance unsupervised 3D DLIR and enable it. The module uses the results of initial registration to generate a co-attention map that describes remaining registration errors spatially and feedback this to the DLIR to minimize such errors and improve self-supervision. We show that FBA improves a range of promising 3D DLIR designs, including networks with and without transformer enhancements, and that it can be applied to both fetal and adult 3D echo, suggesting that it can be widely and flexibly applied. We further find that the optimal 3D DLIR configuration is when FBA is combined with a spatial transformer and DLIR backbone modified with spatial and channel attention, which outperforms existing 3D DLIR approaches. FBA's good performance suggests that spatial attention is a good way to enable scaling up from 2D DLIR to 3D and that a focus on the quality of the image after registration warping is a good way to enhance DLIR performance.

Index Terms—3D+time echocardiography, Adult and fetal echocardiography registration, Cardiac motion estimation, Feedback attention, and Self-supervised DLIR.

I. INTRODUCTION

HEART diseases are a top reason for mortality, and echocardiography is the primary tool for cardiac health assessment. Cardiac motion estimation from echo is a good

means of extracting myocardial strains, cardiac chamber volume changes, and other motion features, which are important indicators of the heart's health and disease severity [1], [2] that inform patient management and intervention strategies.

To date, echo measurements of motion have limited precision in both adult [3] and fetal [4] echo, with high variability across different reports. Deep learning (DL) image registration (DLIR) can improve accuracy and precision to address this [5], [6], and substantially reduce computation time. However, while there are several successful 2D echo DLIR approaches, there have been few studies of 3D echo DLIR (see section II-B), and there remains room for improvement in 3D echo DLIR performance. One possible reason for this is that the increased dimensionality in 3D obscures the anatomy of interest to be a smaller percentage of the image, and a spatial anatomical attention mechanism may be needed.

The current clinical norm is to use 2D echo, but this has many limitations on its ability to describe and visualize the complex 3D anatomy and motion of the heart, which is exacerbated by the increase in complexity during diseases and malformations. Currently, 3D echo is widely available, and studies suggest it can provide accuracy and reproducibility advantages over 2D echo [7]. 3D echo also offers more information to DL algorithms and may improve such algorithms' ability to diagnose and prognose disease. It is thus essential to develop robust 3D echo DLIR algorithms.

Here, we propose a novel feedback attention (FBA) module that can be added to existing self-supervised 3D DLIR networks to improve and enable them. The overview of this approach is shown in Fig. 1. The typical unsupervised DLIR (Fig. 1) minimizes a similarity loss between the fixed image and warped image (moving image warped with registration output). FBA uses the results of the initial registration and generates a co-attention spatial map between the fixed and warped images that describes the locations of remaining errors, which is feedback to the network to improve spatial self-supervision to guide the network towards higher accuracy and better anatomical plausibility (Fig. 1). Here, we demonstrate that FBA improves various 3D DLIR networks and is successful in both adult and fetal echo (such as shown in Fig. 1). We further show that the optimal configuration for 3D echo DLIR consists of FBA together with previously demonstrated transformer-based DLIR. Our specific contributions are:

- We introduce a novel feedback attention (FBA) module

M. K. Hasan was supported by the DTP-studentship funds of the EPSRC (2022–2026) from the Bioengineering department of ICL, UK. G. Yang was supported in part by the ERC IMI (101005122), the H2020 (952172), the MRC (MC/PC/21013), the Royal Society (IEC/NSFC/211235), the NVIDIA Academic Hardware Grant Program, and the UKRI Future Leaders Fellowship (MR/V023799/1).

M. K. Hasan, Y. Luo, and C. H. Yap are with the Bioengineering Department, Imperial College London (ICL), London SW7 2AZ, UK. E-mail: k.hasan22@imperial.ac.uk, y.luo23@imperial.ac.uk, and c.yap@imperial.ac.uk

G. Yang is with the Bioengineering Department and Imperial-X, ICL, London W12 7SL, UK; National Heart and Lung Institute, ICL, London SW7 2AZ, UK. E-mail: g.yang@imperial.ac.uk

[†]Co-last senior authors.

Manuscript received December X, XX; revised August X, XX.

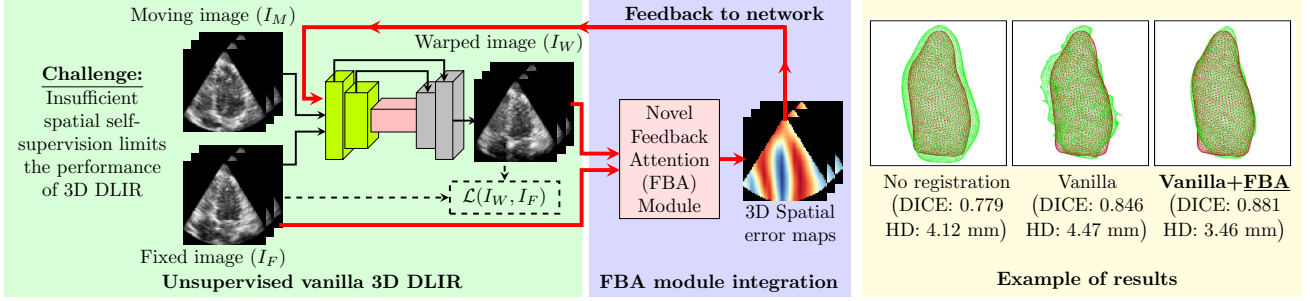


Fig. 1: **Left panel:** Current unsupervised DLIR optimizes similarity ($\mathcal{L}$) with regularizations [5], [8], [9], and has limited performance due to insufficient spatial self-supervision. **Middle panel:** We propose a novel FBA module where a spatial error map of the remaining error of the initial registration is generated, and is feedback to the DLIR network to improve spatial self-supervision. **Right panel:** An example of improved results offered by adding FBA to an unsupervised vanilla DLIR.

for use in unsupervised 3D DLIR networks (Fig. 1). FBA computes a spatial co-attention spatial error map to provide feedback to the DLIR, enhancing its self-supervision. An optimal transportation cost (OTC) loss function is proposed to minimize such errors, where the conversion of the affinity matrices to one devoid of errors is modeled in terms of transportation costs.

- We show that FBA can be flexibly applied to various promising 3D DLIR networks and consistently improves performance when applied, thus demonstrating broad applicability and utility. Such networks include the vanilla DLIR, a spatial transformer DLIR, and DLIR enhanced with spatial and channel attention (SCA) [10].
- We demonstrate that an optimal 3D echo DLIR configuration is coupling FBA with DLIR enhanced with an SCA backbone and a spatial transformer, as these strategies provide non-overlapping benefits, and that the configuration outperforms existing 3D DLIR approaches.
- We demonstrate that this approach has robust 3D DLIR performance in both clinical fetal and adult 3D+time echo images. To date, there is a scarcity of reports of 3D DLIR in actual clinical echo images, as existing reports used animal or synthetic images. We address this here.

II. RELATED WORKS

A. Non-Echocardiographic DLIRs

DLIR can be supervised or unsupervised. Examples of supervised DLIR are in [11] and [12], where ground truths are obtained from segmentations or non-DL registration. While supervised DLIRs are easy to train, obtaining good ground truths is challenging, as non-DL methods used for ground truths are often imperfect. Synthetic images generated based on known fields have also been used for supervision [6], but there are differences between synthetic and clinical images.

Unsupervised DLIRs have the advantage of not requiring ground truths, and by demonstrating results in actual clinical images, they instill confidence. However, additional and often complex constraints are needed for good performance [8], [9], [13]. Typically, unsupervised DLIRs are trained by enforcing a match between the fixed and warped moving images. A particularly well-received algorithm is VoxelMorph [5], which

adds a regularization of minimizing dice-loss between the warped and fixed segmentation labels, thus using anatomy information to guide the DLIR.

VoxelMorph was originally developed for brain MRIs but has been successfully applied to cardiac MRIs [14]. It has been the basis for several DLIR advances. For example, Hasan et al. [8] combined VoxelMorph, which offers pixel-wise constraints, with a shape encoder, which gives a global anatomy representation. Further, the authors in [9] combined CycleGAN and VoxelMorph to achieve cycle consistency using bidirectional deformation fields. A further strategy to enhance DLIR is multi-resolution training, such as that employed by the deep Laplacian pyramid image registration network (LapIRN), proposed in [13] and demonstrated for brain MRIs.

B. Echocardiographic DLIRs

Several DLIR methods have been developed for DLIR in 2D echo. Ostvik et al. [15] proposed the PWC-Net based on a feature pyramid extractor, supervised by an optical flow estimator, and showed that the resulting myocardial strains compared well with truths and current commercial echo machine quantifications. Wei et al. [16] proposed the CLAS algorithm, where a 2D+time UNet performed registration and segmentation concurrently for the entire cardiac cycle and showed that results were improved by constraining consistency between the two tasks. Wang et al. [17] proposed a patch-based MLP and transformer network that utilized a cross-feature block for matching features from moving and fixed images and demonstrated the potential of non-CNN networks.

However, there have been few reports on DLIR in 3D echo. Among them, Lu et al. [18] proposed an MLP-based DLIR supervised by a synthetic dataset that is regularized by biomechanical constraints of tissue incompressibility and cyclic motion of the heart. Further, Ahn et al. [6] proposed a co-attention spatial transformer, leveraging the UNet-based backbone and co-attention techniques [19] in 3D porcine and synthetic images. However, from the results reported, such as DICE indices, there appears to be room for improvement. Here, we aim to extend and improve these state-of-the-art 3D DLIR efforts and implement them in clinical images.

Further, there have been no 3D fetal echo DLIR studies. We strive to address this gap as well. As there are fundamental

differences between fetal and adult echo features, such as the small size of the fetal heart and the poor out-of-plane resolution and image contrast in fetal echo, DLIR methods should ideally be demonstrated for both adult and fetal echo.

C. Overview of attention

Recent literature has demonstrated that attention mechanisms can produce accuracy benefits above traditional DLIRs. For example, a spatial and channel attention (SCA) approach was introduced in [10] to instruct the network on which features and locations to prioritize. Further, several authors successfully employed temporal attention [6], [19]–[21]. This included investigators of non-medical 2D natural videos, Lu et al. [19], who used co-attention to ensure that the primary object frequently and distinguishably appears in all the time frames; and Wang et al. [20], who used an attention-based model that fused the first, previous, and current frames so that discriminative target features are learned and drifting inaccuracies of propagation schemes are reduced. For medical imaging, Ahn et al. [21] used temporal co-attention to assist with left ventricular segmentation in 3D echo, extracting correlated spatiotemporal features in temporal sequences. Further, authors in [6] used a spatiotemporal co-attention [19] for cardiac motion tracking in 3D echo. Unlike these works, our proposed FBA module generates co-attention maps of remaining errors for feedback to the network to improve it.

III. METHODOLOGY AND MATERIALS

A. Proposed feedback DLIR

Fig. 2 depicts the best configuration that we can find with our FBA DLIR, which combines FBA with existing transformer-based DLIR. Removing and/or replacing various components of this amalgamated configuration will provide DLIR networks of various designs. Through ablation studies, we investigate the effectiveness of adding our novel FBA module to various 3D DLIR networks.

This FBA DLIR pipeline consists of 3 network blocks. Blocks **A1** and **A2** are identical Siamese blocks with shared weights, which are SCA-based backbone DLIR blocks that compute deformation fields. **A1** outputs the initial deformation field, φ_i , while **A2** outputs the final deformation field, φ_f . Two spatial co-attention blocks (**B** and **C**) modify the input images going into blocks **A1** and **A2** with a co-attention map derived from the embedded features of a pair of images. Block **B** generates the co-attention map between the original fixed and moving images, which is applied to both images before registration in block **A1**. This co-attention map emphasizes spatial locales that contain descriptions of the heart motions to enhance the registration and is motivated by [6]. Block **C** is our novel feedback module. It generates the co-attention map between the warped (moving image deformed by initial registration output) and fixed images, and the map is then applied to the fixed image before input into block **A2**. Here, the co-attention map is derived from two images that are ideally similar and thus describes spatial locales of dissimilarities and initial registration errors. Since block **C** uses the outputs of initial registration to produce a better registration, it serves as

a form of feedback mechanism to the registration process to guide the training process, thus its name. When adding FBA to a DLIR network, one has the choice of adding only block **C**, where only the fixed image is attention-modified in the final DLIR, or adding both block **B** and **C**, where both the fixed and moving images are attention modified. Further details of blocks **A**, **B**, and **C** are provided below.

1) *Unsupervised DLIR networks (block A1 and A2)*: Blocks **A1** and **A2** are the backbone DLIR ($\mathcal{F}_{\mathcal{CNN}}$) in Fig. 2, consisting of three encoding and decoding convolutional blocks according to the UNet structure, with 128, 256, and 512 filters in each block. It predicts a deformation field $\varphi(x)$ from a pair of N -dimensional input volumes, $\varphi(x) = \mathcal{F}_{\mathcal{CNN}}(I_F, I_M; \Theta)$; where Θ is training parameters and $\varphi(x) : \mathcal{R}^N \mapsto \mathcal{R}^N$. This configuration is named “Vanilla DLIR”. In our “SCA DLIR” and “Feedback DLIR” approaches, we propose to couple the SCA module [10] to the convolutional and deconvolutional blocks of the backbone DLIR. The SCA module follows the outputs of those two blocks to capture enriched inter-dependent anatomical regions and channel features. To train $\mathcal{F}_{\mathcal{CNN}}$ (vanilla or SCA DLIR), equation (1) is optimized.

$$\hat{\Theta} = \max_{\Theta} \{ \mathcal{L}_{img}(I_F, I_M \circ \varphi) + \lambda_r \times \mathcal{L}_r(\varphi) \}, \quad (1)$$

where φ is the deformation field and $\mathcal{L}_{img}([0, 1])$ is an image-based similarity metric, which can be normalized cross-correlation (NCC) or mean square error (MSE). Here, we choose to use MSE [5]. $\mathcal{L}_r$ is a smoothness regularization [5], and $\lambda_r \geq 0$ is the smoothness weight.

2) *FBA self-supervision module (block C) and Spatial attention module (block B)*: For the given fixed and moving images ($I_F, I_M \in \mathbb{R}^{H \times W \times D}$), DLIR aims to find a φ to warp the I_M into $I_W = I_M \circ \varphi \in \mathbb{R}^{H \times W \times D}$ with the best fit to I_F , where H , W , and D indicate the height, width, and depth of the images. In blocks **B** and **C**, 3D co-attention maps based on [19] are computed to emphasize specific anatomical locations. The co-attention maps are computed via feature embedding, which enables the DLIR to understand the association between the two input images in a global feature-embedded space. The same embedding network, $\mathcal{E}$ (Fig. 2), is used throughout the algorithm, based on [22]. It consists of a series of 3D convolutions with 64, 128, and 256 filters and an atrous spatial pyramid pooling module using parallel dilated convolutions with different rates (6, 12, and 18) to capture multi-scale contextual information. $\mathcal{E}$ returns the embedded features map, $\Psi = \mathcal{E}(I) \in \mathbb{R}^{C(=256) \times h \times w \times d}$, where $h, w, d = H/4, W/4, D/4$.

The $\mathcal{E}$ is trainable, and its learning is partly constrained by the agreement maximization between the fixed and moving embedded representations, Ψ_F and Ψ_M . This is motivated by the self-supervised embedding learning proposed in [23], [24]. The constraint minimizes the redundancy between the components of the embedded representations and, consequently, returns a better co-attention map that is minimally sensitive to the differences in the two images caused by the heart’s movement. In our FBA DLIR, we also propose to use the normalized cross-correlation $\mathcal{NCC}$ matrix between Ψ_F and Ψ_M as an objective function [23], as shown in (2), and to

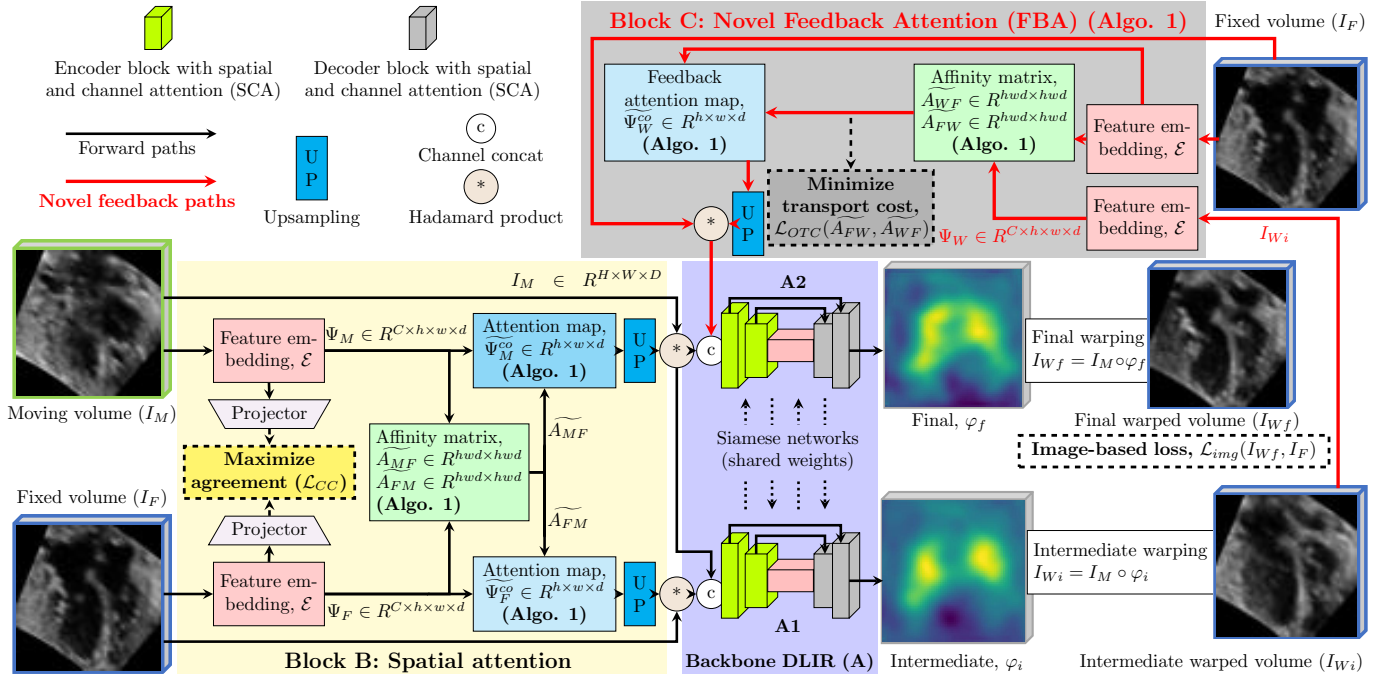


Fig. 2: The proposed self-supervised DLIR pipeline using our novel FBA (block C) module, incorporating Siamese backbone DLIR blocks (block A1 and A2) and spatial attention (block B). Feature embedding blocks $\mathcal{E}$ s are Siamese encoders. All blocks A, B, and C are trainable. Our FBA module (Algorithm 1) in block C facilitates the DLIR's self-supervision.

Algorithm 1: Estimation of FBA (block C) and spatial co-attention (block B).

Input: An image pair $\{I_F, I_M\}$ or $\{I_F, I_W\} \in \mathbb{R}^{H \times W \times D}$

Output: Spatial FBA ($\xi_W \in \mathbb{R}^{H \times W \times D}$) using $\{I_F, I_W\}$ or spatial co-attention ($\xi_{F/M} \in \mathbb{R}^{H \times W \times D}$) using $\{I_F, I_M\}$

- 1 Extract embedded semantic feature maps as: $\Psi_F = \mathcal{E}(I_F)$, $\Psi_M = \mathcal{E}(I_M)$, and $\Psi_W = \mathcal{E}(I_W) \in \mathbb{R}^{C \times h \times w \times d}$, where $\mathcal{E}$ is a feature embedding Siamese encoder and C is the channel numbers
- 2 Compute affinity matrix (A) to mine the correlations between Ψ_F & Ψ_M or Ψ_F & Ψ_W in their feature embedding space by flattening those maps ($\Psi^l \in \mathbb{R}^{C \times hwd}$) and using a trainable linear transformation, $\Psi^l = \Psi^l W_l^T + b$, where W_l and b are the linear transformation weight matrix and bias

$$A_{FM} = \Psi_F^l \cdot W_{fm} \cdot \Psi_M^l \quad \text{or} \quad A_{MF} = \Psi_M^l \cdot W_{mf} \cdot \Psi_F^l \in \mathbb{R}^{hwd \times hwd},$$

$$A_{FW} = \Psi_F^l \cdot W_{fw} \cdot \Psi_W^l \quad \text{or} \quad A_{WF} = \Psi_W^l \cdot W_{wf} \cdot \Psi_F^l \in \mathbb{R}^{hwd \times hwd},$$

where $W \in \mathbb{R}^{C \times C}$ is a diagonal weight matrix [19].

- 3 Normalize all the affinity matrices using a softmax activation (σ_1) as $\tilde{A} = \sigma_1(A)$, which reflects the relevance of each feature ($k \in \{1, 2, \dots, hwd\}$) in Ψ_F to the k^{th} feature in Ψ_M or Ψ_W
 - 4 Estimate co-attention enhanced attention feature map as: $\Psi_F^{co} = \Psi_F \cdot \tilde{A}_{FM} \in \mathbb{R}^{C \times hwd}$ (for fixed image) and $\Psi_M^{co} = \Psi_M \cdot \tilde{A}_{MF} \in \mathbb{R}^{C \times hwd}$ (for moving image), and $\Psi_W^{co} = \Psi_F \cdot \tilde{A}_{FW} \in \mathbb{R}^{C \times hwd}$ (for feedback attention).
 - 5 Apply post-processing to those enhanced feature maps (unflattened $\in \mathbb{R}^{C \times h \times w \times d}$) using a convolutional kernel (γ), bias (Ω), and sigmoid activation (σ_2) with an output channel of 1: $\tilde{\Psi}^{co} = \sigma_2(\gamma \Psi^{co} + \Omega) \in [0, 1]$, where $\tilde{\Psi}^{co} \in \mathbb{R}^{h \times w \times d}$
 - 6 Upsampling the refined attention maps using a trilinear interpolation (Υ) provides: $\xi_F = \Upsilon(\tilde{\Psi}_F^{co}) \in \mathbb{R}^{H \times W \times D}$, $\xi_M = \Upsilon(\tilde{\Psi}_M^{co}) \in \mathbb{R}^{H \times W \times D}$, and $\xi_W = \Upsilon(\tilde{\Psi}_W^{co}) \in \mathbb{R}^{H \times W \times D}$
-

define a loss function, $\mathcal{L}_{CC}(\Psi_F, \Psi_M)$, as described in (3).

$$\mathcal{NCC}(I, J) = \frac{\sum_{p \in x, y, z}^{H, W, D} (I(p) - \mu_I)(J(p) - \mu_J)}{\sqrt{\sum_{p \in x, y, z}^{H, W, D} (I(p) - \mu_I)^2} \sqrt{\sum_{p \in x, y, z}^{H, W, D} (J(p) - \mu_J)^2}}, \quad (2)$$

$$\mathcal{L}_{CC}(\Psi_F, \Psi_M) = \sum_c \left(1 - \mathcal{NCC}(z_{b,c}^{\Psi_F}, z_{b,c}^{\Psi_M})\right)^2 + \gamma \sum_c \sum_{c \neq k} \mathcal{NCC}(z_{b,c}^{\Psi_F}, z_{b,k}^{\Psi_M})^2, \quad (3)$$

where $z = g(\cdot)$ is the output of a projection head [24] $g(\cdot)$, as $z^\Psi = g(\Psi)$, where $g(\cdot)$ is a small MLP block with two

hidden layers with 512 neurons. $\mathcal{NCC}(z_{b,c}^{\Psi_F}, z_{b,c}^{\Psi_M})$ is the cross-correlation between the corresponding features in Ψ_F and Ψ_M , while $\mathcal{NCC}(z_{b,c}^{\Psi_F}, z_{b,k}^{\Psi_M})$ is the cross-correlation between non-corresponding features in the two. In essence, $\mathcal{L}_{CC}$ maximizes the correlation of corresponding features and minimizes the correlation of non-corresponding features to minimize the embedded representation redundancy discussed above. γ is a hyperparameter set as 0.005 in our experiments, and b is an index within the batch size.

Algorithm 1 explains the procedures of spatial FBA (ξ_W) and co-attention (ξ_F and ξ_M) estimation. Briefly, this entails obtaining the embedded features (Ψ) of images, computing the affinity matrix (A) of features from the pair of images, normalizing the affinity matrix, and applying it to features of each image, followed by post-processing and up-sampling. The obtained co-attention maps (ξ_F and ξ_M) and the FBA co-attention map (ξ_W) are then applied to the input image volumes to highlight locations of importance.

In the FBA block **C**, a co-attention map between the fixed and warped images (moving images warped with the initial registration) is calculated. With the ideal registration, these two images should be similar, and as such, during the training of FBA, we impose an additional constraint to maximize the similarity between them. We model this loss constraint as minimizing the cost of transporting weights between the fixed-to-warped image affinity matrix and the warped-to-fixed image affinity matrix, such that with minimization, the affinity matrix will be equivalent to one that maximizes the similarity in features between the two images. The optimal transportation cost (OTC) function is defined in (4).

$$OTC(\widetilde{A_{FW}}, \widetilde{A_{WF}}) = \min_{\pi \in \Pi(\widetilde{A_{FW}}, \widetilde{A_{WF}})} \iint d(x, y) \pi(x, y) dx dy, \quad (4)$$

where the cost is defined as the distance ($d(x, y)$) between the index locations x and y in the Affinity matrix. The index locations of the 3D affinity matrix are modeled to be a 3D cartesian grid, with consecutive indices assumed to have a spacing distance of 1. The transportation plan, π , is defined as the quantity of weights to be transported when converting from $\widetilde{A_{WF}}$ to $\widetilde{A_{FW}}$. The OTC loss function is thus in (5).

$$\mathcal{L}_{OTC}(\widetilde{A_{FW}}, \widetilde{A_{WF}}) = \frac{1}{2B} \sum_{b=1}^B \left(\sum_{i,j}^{(H/2^n)^3} \widetilde{A_{FW}}(b, i, j) \odot d(i, j) + \sum_{i,j}^{(H/2^n)^3} \widetilde{A_{WF}}(b, i, j) \odot d(i, j) \right), \quad (5)$$

where i and j are similarly $= 1, 2, 3, \dots, hwd \times hwd$, corresponding to all the entries in the affinity matrix, b is an index in the batch with a batch size of B , and $\odot$ is the element-wise multiplication.

On top of this, two further loss constraints are imposed on the attention map and attention-modified images. These constraints create layers of redundancy to ensure easy convergence and accurate results. The first constraint (6) is applied to attention maps to maximize the similarity between the attention map for the fixed image (either ξ_F in spatial attention

block, block **B**, or ξ_W in the feedback block, block **C**) and that for the moving image (ξ_M) warped with registration output (φ). Attention maps are constrained using (6) via normalized cross-correlation according to (2).

$$\mathcal{L}_{attention}(\xi_M, \xi_F, \xi_W, \varphi_i, \varphi_f) = \mathcal{NCC}(\xi_W, \xi_M \circ \varphi_f) + \mathcal{NCC}(\xi_F, \xi_M \circ \varphi_i), \quad (6)$$

where $\mathcal{NCC}(\xi_W, \xi_M \circ \varphi_f)([0, 1])$ and $\mathcal{NCC}(\xi_F, \xi_M \circ \varphi_i)([0, 1])$ are the spatial feedback and spatial co-attention terms in the attention loss ($\mathcal{L}_{attention}$), respectively.

The second constraint is applied to attention-modified images. In blocks **B** and **C**, co-attention maps are applied to the original images before inputted into the DLIR blocks (**A1** and **A2**): $I'_F = I_F \times \xi_F$, $I'_M = I_M \times \xi_M$, $I''_F = I_F \times \xi_W \in \mathbb{R}^{H \times W \times D}$. An MSE loss constraint ($\mathcal{M}([0, 1])$) is imposed to maximize the similarity between the attention-modified fixed (or feedback) image and the attention-modified warped image, according to (7).

$$\mathcal{M}(I, J, \varphi) = \frac{1}{H \times W \times D} \sum_{p \in x, y, z}^{H, W, D} (I(p) - J(p) \circ \varphi)^2. \\ \mathcal{L}_{refined_img}(I'_M, I'_F, I''_F, \varphi_i, \varphi_f) = \mathcal{M}(I''_F, I'_M, \varphi_f) + \mathcal{M}(I'_F, I'_M, \varphi_i), \quad (7)$$

where $\mathcal{M}(I'_F, I'_M, \varphi_i)$ enforces the co-attention in the initial registration, while $\mathcal{M}(I''_F, I'_M, \varphi_f)$ provides self-supervision as feedback.

3) The optimal FBA DLIR: The final loss function for our optimal FBA DLIR configuration in Fig. 2 is defined in (8), where the terms $\mathcal{NCC}(\xi_W, \xi_M \circ \varphi_f)$ in (6) and $\mathcal{M}(I''_F, I'_M, \varphi_f)$ in (7) incorporate the feedback co-attention.

$$\hat{\Theta} = \max_{\Theta} \left\{ \lambda_1 \mathcal{L}_{img}(I_F, I_M \circ \varphi_f) + \lambda_2 \times \mathcal{L}_r(\varphi_f) + \lambda_3 \times \mathcal{L}_{CC}(\Psi_F, \Psi_M) + \lambda_4 \times \mathcal{L}_{OTC}(\widetilde{A_{WF}}, \widetilde{A_{FW}}) + \lambda_5 \times \mathcal{L}_{attention}(\xi_M, \xi_F, \xi_W, \varphi_i, \varphi_f) + \lambda_6 \times \mathcal{L}_{refined_img}(I'_M, I'_F, I''_F, \varphi_i, \varphi_f) \right\}, \quad (8)$$

where λ_{1-6} are the weighting factors. Depending on the individual loss terms values as discussed earlier, those weighting factors are set $\lambda_1=10$, $\lambda_2=10$, $\lambda_3=1e-4$, $\lambda_4=1e-5$, $\lambda_5=1$, and $\lambda_6=1$ by trials and errors. Algorithm 2 shows PyTorch-style pseudocode of our feedback DLIR's implementation.

B. Datasets and Implementation

1) Public adult 3D echocardiography: We utilize the MITEA dataset for training and testing, a 3D adult echo comprising 134 patients with 268 sample acquisitions at the end-diastolic (ED) and end-systolic (ES) [25], which gives 536 3D image volumes (details in [25]). The volumes are provided with anatomical segmentation masks. The sample distribution for training-validation/testing of seven different classes is as follows: 328 (304/24) healthy, 56 (48/8) with left ventricular hypertrophy, 48 (36/12) with cardiac amyloidosis, 40 (32/8) with aortic regurgitation, 32 (28/4) with hypertrophic cardiomyopathy, 24 (20/4) with dilated cardiomyopathy, and 8 (4/4) with heart transplant recipients.

Algorithm 2: PyTorch-style pseudocode of FBA DLIR

```

1 for  $I_F, I_M$  in loader: do
2   # Forward pass
3   # Algorithm 1 for BLOCK B
4    $\Psi_F, \Psi_M, \rightarrow, \rightarrow, \xi_F, \xi_M = \text{Attention}(I_F, I_M)$ 
5   # Refine fixed and moving volumes
6    $I'_F = I_F \times \xi_F$  and  $I'_M = I_M \times \xi_M \in \mathbb{R}^{H \times W \times D}$ 
7   # Train our SCA-DLIR (BLOCK A1)
8    $\varphi_i = \mathcal{F}_{CNN}(I'_F, I'_M)$  and  $I_{Wi} = I_M \circ \varphi_i$ 
9   # Algorithm 1 for FBA (BLOCK C)
10   $\rightarrow, \rightarrow, \widetilde{A}_{WF}, \widetilde{A}_{FW}, \rightarrow, \xi_W = \text{Attention}(I_F, I_{Wi})$ 
11  # Refine fixed volumes w.r.t initial warped ( $I_{Wi}$ )
12   $I''_F = I_F \times \xi_W \in \mathbb{R}^{H \times W \times D}$ 
13  # Train our SCA-DLIR (BLOCK A2)
14   $\varphi_f = \mathcal{F}_{CNN}(I''_F, I'_M)$  and  $I_{Wf} = I_M \circ \varphi_f$ 
15  # Loss estimation
16   $\mathcal{L}_{img}(I_F, I_M \circ \varphi_f)$  # Image-based loss
17   $\mathcal{L}_r(\varphi_f)$  # Smoothness regularizer
18   $\mathcal{L}_{CC}(\Psi_F, \Psi_M)$  # Feature embedding loss
19   $\mathcal{L}_{OTC}(\widetilde{A}_{WF}, \widetilde{A}_{FW})$  # Affinity matrix loss
20   $\mathcal{L}_{attention}(\xi_M, \xi_F, \xi_W, \varphi_i, \varphi_f)$  # Attention loss
21   $\mathcal{L}_{refined\_img}(I'_M, I'_F, I''_F, \varphi_i, \varphi_f)$  # Refined loss
22   $loss \leftarrow \text{Apply (8) for overall loss for FBA DLIR}$ 
23  # Backward pass
24   $optimizer.zero\_grad()$ 
25   $loss.backward()$ 
26   $optimizer.step()$  # Update model parameters

```

2) *Private fetal 3D echocardiography:* We utilize a private 4D (3D+time) fetal echo dataset comprising 26 fetal patients with 25 to 40 time frames, resulting in a total of 983 3D volumes with an in-plane resolution of $0.95\mu\text{m} \times 0.90\mu\text{m}$ (see details in [26]). Ethics approval was obtained under Kepler University Hospital IRB protocol 1009/2017 and the NUS DSRB protocol 2014/00056. The anatomical regions are manually labeled (details in [8]). A total of 156 fixed and moving volumes were selected from 983 volumes by various combinations of the time frames near the ES and those near the ED, ensuring a considerable deformation between fixed and moving volumes. The sample distribution for training/validation and testing is 138 and 18 volumes.

3) *Training protocol and evaluation:* DLIRs are implemented using PyTorch and MONAI. The batch size and epochs are eight and three hundred, utilizing an Adam optimizer with a learning rate of $1e-5$. For computational efficiency, before testing and training, all the images and masks are resampled to $64 \times 64 \times 64$ using bicubic interpolation for intensity images and nearest-neighbor interpolation for the masks. The Ubuntu machine used for the experiments is equipped with 4 NVIDIA® GeForce RTX® 3090Ti PCIe 3.0 GPUs with 24GB GDDR6X memory with Intel® Xeon® W-2295 x18 @ 3.00GHz CPU and 128 GB of memory.

Results are evaluated using dice similarity coefficient (DSC) for overlapping precision, Hausdorff distance (HD) for border irregularity, and the non-positive rate of the Jacobian

determinant ($|\mathbf{J}_{\varphi_f}|$) to guarantee the diffeomorphism without orientation flipping and local singularity (degeneration) in which the surrounding space of a voxel is warped, as in (9).

$$\begin{aligned}
DSC(M_F^C, M_W^C) &= \frac{2 \cdot |M_F^C \cap (M_M^C \circ \varphi_f)|}{|M_F^C| + |M_M^C \circ \varphi_f|}, \forall k (= 2) \\
HD(M_F^C, M_W^C) &= \\
\max \left\{ \max_{a \in M_F^C} \mathcal{D}(a, M_M^C \circ \varphi_f), \max_{b \in M_M^C \circ \varphi_f} \mathcal{D}(b, M_F^C) \right\}, \forall k (= 2) \\
\text{where } \mathcal{D}(a, M_M^C \circ \varphi_f) &= \min_{b \in M_M^C \circ \varphi_f} \|b - a\|_2 \\
\mathcal{D}(b, M_F^C) &= \min_{a \in M_F^C} \|a - b\|_2, \\
|\mathbf{J}_{\varphi_f}| \leq 0 \Big|_{\text{rates}} &= \frac{\sum(|\nabla \varphi_f| \leq 0)}{V}, \quad (9)
\end{aligned}$$

where ($a \in M_F^C, b \in M_M^C \circ \varphi_f$) are the discrete points on the contours M_F^C and $M_M^C \circ \varphi_f$, respectively. $\mathcal{C}$ is the number of anatomic organs. V is the total number of elements in φ_f .

IV. EXPERIMENTAL RESULTS

A. Effects of FBA on various DLIR networks

We first evaluate if FBA can improve existing 3D echo DLIRs, experimenting with three DLIRs: vanilla, SCA, and co-attention [6] DLIRs. The vanilla DLIR and SCA DLIR retain only block A1 in Fig. 2. In vanilla DLIR, A1 is not SCA enhanced [10]. Co-Attention DLIR retains only blocks A1 and B. Results are shown in Table I. Here, DSC results generally vary across different methods for the endocardial boundary, but this is not so for the epicardial boundary. This is because the epicardial boundary does not move as much, and even without registration, the endocardial and epicardial DSCs are 0.621 ± 0.108 and 0.826 ± 0.047 .

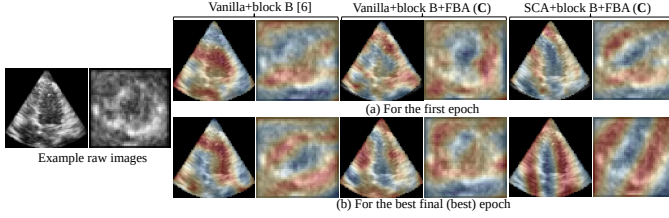
When FBA (block C) is added to vanilla DLIR, vanilla+co-attention DLIR, SCA DLIR, or SCA+co-attention DLIR, improvements to DSC and HD of both endo- and epicardium are generally observable (see Table I), with the exception of HD for the endocardium when FBA is added to vanilla+co-attention DLIR. All improvements are found to be statistically significant ($p \ll 0.05$). Adding FBA also leads to improvements in $|\mathbf{J}_{\varphi_f}|$ for vanilla DLIR and SCA DLIR, suggesting better spatial consistency. However, for networks with co-attention, $|\mathbf{J}_{\varphi_f}|$ is already excellent, and adding FBA worsens it slightly but not significantly. Our results thus show that FBA can be added to a range of DLIR designs and consistently improves those unsupervised DLIRs by providing spatial self-supervision.

B. Analysis of FBA co-attention map

The attention maps overlaid onto the corresponding intensity images are visualized in Fig. 3, where the hot (red) color represents greater spatial emphasis. The first and best epochs are shown in figures (a) and (b), respectively, to show the epoch convergence of the maps. Generally, in the first epoch, attention maps do not coincide with anatomic structures well, with much emphasis placed on locations outside the myocardium. However, in the best epoch, most of the attention

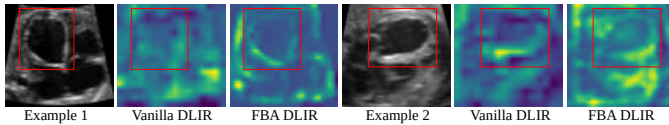
TABLE 1: The quantitative registration results on the publicly available adult MITEA dataset [25] for different DLIR configurations with and without our novel FBA. Bold-underlined fonts indicate the best metrics.

DLIR configurations	DSC ($\uparrow$)		HD (mm) ($\downarrow$)		$ J\varphi_f \leq 0$ rates ($\downarrow$)
	Endocardial	Epicardial	Endocardial	Epicardial	
Vanilla DLIR	0.749 $\pm$ 0.069	0.834 $\pm$ 0.039	7.59 $\pm$ 2.11	7.71 $\pm$ 1.98	11.2 $\pm$ 0.945 %
Vanilla DLIR + FBA	0.767 $\pm$ 0.077	0.845 $\pm$ 0.041	6.09 $\pm$ 1.66	5.63 $\pm$ 1.50	0.004 $\pm$ 0.009 %
Co-Attention DLIR [6]	0.760 $\pm$ 0.085	0.851 $\pm$ 0.042	5.79 $\pm$ 1.80	6.02 $\pm$ 1.99	0.0 $\pm$ 0.0 %
Co-Attention DLIR + FBA	0.785 $\pm$ 0.072	0.852 $\pm$ 0.043	5.93 $\pm$ 1.78	5.66 $\pm$ 1.58	0.025 $\pm$ 0.028 %
SCA DLIR (SCA-enhanced vanilla DLIR)	0.751 $\pm$ 0.073	0.841 $\pm$ 0.042	6.82 $\pm$ 1.93	6.90 $\pm$ 2.15	3.36 $\pm$ 0.520 %
SCA DLIR + FBA	0.783 $\pm$ 0.073	0.864 $\pm$ 0.037	4.89 $\pm$ 1.41	5.56 $\pm$ 1.46	0.041 $\pm$ 0.056 %
SCA-Co-Attention DLIR	0.784 $\pm$ 0.078	0.873 $\pm$ 0.035	4.74 $\pm$ 1.29	4.96 $\pm$ 1.56	0.0 $\pm$ 0.0 %
SCA-Co-Attention DLIR + FBA	0.800 $\pm$ 0.070	0.884 $\pm$ 0.032	4.32 $\pm$ 1.03	4.52 $\pm$ 1.66	0.093 $\pm$ 0.119 %

**Fig. 3:** Attention maps overlaid on the raw images for the MITEA dataset, showing that myocardial structures are more focused in the red region on our FBA DLIR than on others.

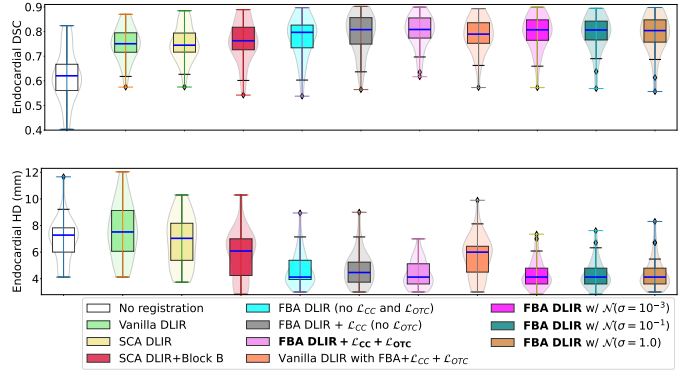
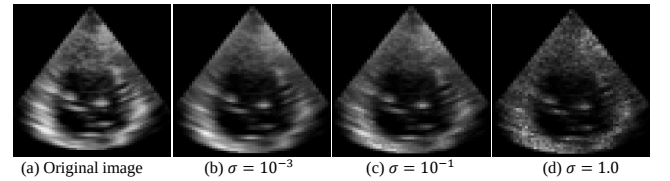
shifted to locations of the myocardium and locations where motions and, thus, errors are larger. Both co-attention DLIR [6] and FBA DLIR utilize co-attention maps, but the visualization demonstrates that adding FBA to co-attention DLIR causes further changes to the co-attention map and conforms even better with anatomic structures. This suggests an enhancement to the spatial self-supervision offered by FBA.

We further visualize the registration feature maps from the second layer of the backbone’s encoder to reveal the improvement in learning of FBA DLIR compared to the vanilla DLIR. As shown in Fig. 4, with FBA DLIR, the feature map coincided with cardiac structures better than vanilla DLIR, again suggesting a successful improvement in spatial emphasis guidance to the network.

**Fig. 4:** Feature maps for vanilla and our FBA DLIRs for the fetal dataset, showing that our FBA retrieves myocardial structural information better than vanilla (see red box).

C. Ablation and Noise Studies in FBA

Next, we perform ablation studies on FBA DLIR to understand the importance of its various components and verify that it is the optimal network for 3D echo DLIR. We observe a progressive improvement of mean DSC and HD when going from vanilla DLIR to SCA DLIR (see Fig. 5), and when blocks **B** (to give co-attention DLIR) and **C** (to give the final FBA DLIR) are sequentially added, but without $\mathcal{L}_{CC}$ and $\mathcal{L}_{OTC}$.

**Fig. 5:** The impact of various integral components in our FBA DLIR (first eight plots) and the impact of different noise levels ($\mathcal{N}(\mu = 0, \sigma)$) on the test volumes compared to the original test volumes (last three plots) for the MITEA dataset.**Fig. 6:** Original test and noise-affected ($\mathcal{N}(\mu, \sigma)$) images at variable levels of σ to show our FBA DLIR’s noise sensitivity.

Subsequently, mean DSC continues to improve when adding $\mathcal{L}_{CC}$ and then $\mathcal{L}_{OTC}$. However, the addition of blocks **B** and **C** increased the spread of DSC results, creating more outliers, and only with imposing $\mathcal{L}_{CC}$ and then $\mathcal{L}_{OTC}$ does the DSC spread improve. With HD, adding SCA and blocks **B** and **C** brought about similar improvements (decrease), further adding $\mathcal{L}_{CC}$ does not improve HD, but adding $\mathcal{L}_{OTC}$ improves the average and spread of HD. Collectively, these results demonstrate that all the proposed components contribute to the excellent performance of the final FBA DLIR approach we proposed, suggesting that it is the optimal configuration.

To evaluate how well our FBA DLIR can handle noise, multiplicative noise is added to the test images ($I_{F/M}^t$) using the formula $I_{F/M}^t = I_{F/M}^t + \mathcal{N}(\mu, \sigma) \times I_{F/M}^t$, where $\mathcal{N}(\mu, \sigma)$ is a uniformly distributed random noise with a mean (μ) and variance (σ [0, 1]). We investigated different variances ($\sigma = \{10^{-3}, 10^{-1}, 1.0\}$) to introduce varying degrees of noise with

TABLE II: The quantitative registration results on the publicly available adult MITEA [25] and in-house multi-demographic fetal echo datasets. Bold-underlined fonts indicate the best metrics for each dataset.

Datasets	Methods	DSC ($\uparrow$)		HD (mm) ($\downarrow$)		$ J_{\varphi_f} \leq 0 \big _{rates}$ ($\downarrow$)	Inference time ($\downarrow$)
		Endocardial	Epicardial	Endocardial	Epicardial		
3D Adult Echo [25] (Public dataset)	ANTs (SyN) [27]	0.655 $\pm$ 0.113	0.851 $\pm$ 0.041	6.28 $\pm$ 1.42	5.90 $\pm$ 1.93	0.794 $\pm$ 0.790 %	$\approx$ 44888 ms
	Elastix [28]	0.657 $\pm$ 0.102	0.853 $\pm$ 0.037	6.56 $\pm$ 1.73	6.84 $\pm$ 1.92	3.61 $\pm$ 2.02 %	$\approx$ 65333 ms
	VoxelMorph [5]	0.749 $\pm$ 0.070	0.848 $\pm$ 0.041	6.80 $\pm$ 1.91	6.66 $\pm$ 1.84	5.46 $\pm$ 0.496 %	$\approx$ 174 ms
	Co-Attention DLIR [6]	0.760 $\pm$ 0.085	0.851 $\pm$ 0.042	5.79 $\pm$ 1.80	6.02 $\pm$ 1.99	0.0 $\pm$ 0.0 %	$\approx$ 251 ms
	LapIRN [13]	0.769 $\pm$ 0.067	0.847 $\pm$ 0.044	6.17 $\pm$ 1.69	6.28 $\pm$ 1.84	2.14 $\pm$ 0.312 %	$\approx$ 171 ms
	SwinUNetR DLIR [29]	0.753 $\pm$ 0.082	0.844 $\pm$ 0.041	5.84 $\pm$ 1.58	6.17 $\pm$ 1.83	0.018 $\pm$ 0.015 %	$\approx$ 250 ms
	UNetR DLIR [30]	0.772 $\pm$ 0.072	0.844 $\pm$ 0.042	5.60 $\pm$ 1.50	6.07 $\pm$ 1.67	0.027 $\pm$ 0.024 %	$\approx$ 201 ms
	Our Feedback DLIR	0.800 $\pm$ 0.070	0.884 $\pm$ 0.032	4.32 $\pm$ 1.03	4.52 $\pm$ 1.66	0.093 $\pm$ 0.119 %	$\approx$ 770 ms
3D Fetal Echo (In-house private dataset)	ANTs (SyN) [27]	0.877 $\pm$ 0.024	0.928 $\pm$ 0.011	4.67 $\pm$ 1.95	4.58 $\pm$ 1.19	4.33 $\pm$ 2.81 %	$\approx$ 43985 ms
	Elastix [28]	0.896 $\pm$ 0.018	0.936 $\pm$ 0.008	3.08 $\pm$ 0.60	3.87 $\pm$ 1.63	0.69 $\pm$ 0.93 %	$\approx$ 66120 ms
	VoxelMorph [5]	0.900 $\pm$ 0.014	0.938 $\pm$ 0.010	3.38 $\pm$ 0.87	3.24 $\pm$ 0.67	0.0003 $\pm$ 0.0002 %	$\approx$ 170 ms
	Co-Attention DLIR [6]	0.910 $\pm$ 0.026	0.934 $\pm$ 0.017	3.07 $\pm$ 0.96	3.50 $\pm$ 0.60	0.0 $\pm$ 0.0 %	$\approx$ 250 ms
	LapIRN [13]	0.907 $\pm$ 0.016	0.936 $\pm$ 0.014	3.19 $\pm$ 0.80	3.15 $\pm$ 0.25	0.0 $\pm$ 0.0 %	$\approx$ 178 ms
	SwinUNetR DLIR [29]	0.912 $\pm$ 0.018	0.937 $\pm$ 0.007	3.19 $\pm$ 0.90	3.42 $\pm$ 0.78	0.0 $\pm$ 0.0 %	$\approx$ 241 ms
	UNetR DLIR [30]	0.897 $\pm$ 0.013	0.939 $\pm$ 0.011	3.25 $\pm$ 0.78	3.34 $\pm$ 0.86	0.0 $\pm$ 0.0 %	$\approx$ 212 ms
	Our Feedback DLIR	0.920 $\pm$ 0.014	0.944 $\pm$ 0.011	3.10 $\pm$ 0.75	3.15 $\pm$ 0.49	0.0 $\pm$ 0.0 %	$\approx$ 791 ms

a fixed mean of $\mu = 0$ (see Fig. 6). Up to the variance of 10^{-1} , the FBA DLIR returns in similar average DSC and HD with a little increase in the interquartile range of the box-violin plots (see Fig. 5). We further test a variance of 1.0, which results in very high noise in the test images (see Fig. 6 (d)). However, our proposed FBA DLIR performs similarly well, except that the variance of DSC and HD increases, as there are more outliers with poorer performance. This demonstrates that the FBA DLIR is sufficiently robust to deal with noise.

D. The optimal FBA 3D Echo DLIR

Our ablation studies demonstrate that the optimal FBA DLIR is when FBA is combined with SCA and co-attention, embodied by all blocks in Fig. 2. Here, we compare it to various alternative DLIRs to show that it provides improvements. This included two well-received non-DL methods, ANTs (SyN) [27] and Elastix [28], VoxelMorph [5], co-attention DLIR [6], LapIRN [13], SwinUNetR [29] and UNetR [30]. Results are summarized in Table II and Fig. 7. It is noted that DSC results in Table II for the fetal echo are higher than adult echo because several individuals in the fetal dataset have a disease that reduces left ventricular myocardial motion.

Endocardial DSC results generally show that non-DL methods, ANTs (SyN) and Elastix, do not work as well as DLIR approaches. Their endocardial DSC for adults is significantly lower than ($p \ll 0.05$) the baseline vanilla DLIR, which is generally similar to or lower than other DLIRs. However, results for average HD, $|J_{\varphi_f}|$, and mesh visualization in Fig. 7 indicate that non-DL techniques are better regularized and are smoother than vanilla DLIR. Our vanilla DLIR utilizes Transform Bending Energy for smoothness regularization, but stronger smoothness comes at the expense of DSC accuracy. In our implementation, we optimize for maximum DSC and consequently observe substantial border irregularities. Further, non-DL approaches take substantially longer to process than all DLIRs. Vanilla DLIR performance is generally not satisfactory; this is unsurprising as it does not have enough spatial, anatomical, and feature guidance.

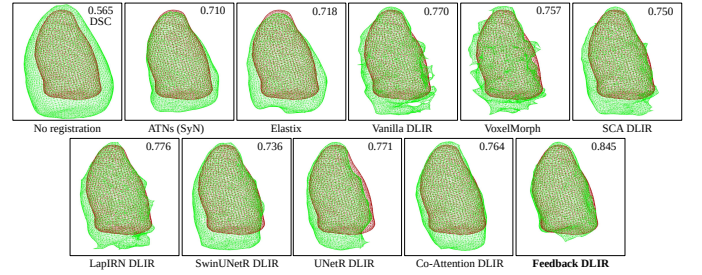


Fig. 7: The endocardial 3D mesh of the warped moving (green color) w.r.t. fixed volumes (red color), the top-left sample is for no registration, and the remaining are from the different methods on the MITEA dataset.

The VoxelMorph incorporates a DSC-based auxiliary loss to vanilla DLIR. This improves HD and $|J_{\varphi_f}|$ for both datasets while maintaining similar inference times and DSC accuracy. Fig. 7 shows an example of VoxelMorph's outputs, showing that substantial irregular spikes remain in the endocardial surface. Interestingly, our observation here that VoxelMorph does not improve DSC accuracy in 3D echo contrasted with our previous experience in 2D echo where it provided improvements [8]. This shows that successful strategies for enhancing 2D DLIR may not be effective in 3D DLIR.

The co-attention DLIR (vanilla DLIR with **B**), as advocated in [6], improves DSC from vanilla DLIR, raising the average DSC for adult echo and reducing the variance of endocardial DSC for fetal echo. It also provides significant ($p \ll 0.05$) progress in HD and $|J_{\varphi_f}|$. These results corroborate with reports in [6]. Interestingly, co-attention reduces $|J_{\varphi_f}|$ to 0, and a visual observation of the resulting warped endocardial boundary (Fig. 7) shows smooth surfaces. This DLIR thus naturally ensures spatial smoothness, which is also reported in [6].

Two other self-attention-based DLIRs, SwinTrans [29] and UNetR [30], which use multi-head self-attention blocks in the encoder, are tested. Results show that they perform similarly

to co-attention DLIR, with DSC, HD, and $|\mathbf{J}_{\varphi_f}|$ close to those of co-attention DLIR, although a dip in performance in endocardial DSC is observed for UNetR for fetal echo. Further, the non-attention-based, multi-scale LapIRN is tested and found to perform similarly to SwinTrans and UNetR. However, visual observation of endocardial boundary mesh (Fig. 7) shows that attention-based DLIRs typically have smoother anatomical surfaces.

In comparison, our optimal FBA DLIR provides significant ($p \ll 0.05$) improvements to epicardial and endocardial DSC beyond all other methods tested in Table II. HD is significantly reduced from all other methods for adult echo and is statistically insignificantly different from the best methods for fetal echo. Further, $|\mathbf{J}_{\varphi_f}|$ is statistically insignificantly different from 0 for adult echo and is 0 for fetal echo. Fig. 7 shows that FBA results in warped volumes with few endocardial and myocardial border irregularities and spikes and gives a very good visual fit at the apex and base of the heart. Further, Table II shows that all methods perform differently across the two datasets, while our FBA DLIR provides consistent results regarding the DSC and HD, revealing FBA DLIR's robustness and genericity. Results thus indicate that FBA can improve on existing reported 3D echo DLIRs and has the robustness and genericity to do well in both adult and fetal echo.

E. Performance in varied disease classifications

The adult echo dataset contains cases of different diseases. We perform subgroup analysis to understand the relative performance of the various DLIRs investigated in various disease subgroups. Interestingly, as shown in Fig. 8, results show varied performance across different cardiac conditions. Overall, results show that challenges can be detected for hypertrophic hearts, where the results are typically worse across all methods. In contrast, the results are better with cardiac amyloidosis and aortic regurgitation. This is most likely due to the extent of cardiac motion and deformation, as indicated by the no-registration DSC. During LV hypertrophy, motions are larger, and no-registration DSC is lower, and as such, the result is generally lower, but during cardiac amyloidosis, myocardial strains are generally compromised [31], and the no-registration DCS is higher, and as such, registration is easier, and the performances of various DLIRs are generally higher. Across the different conditions, FBA DLIR consistently performed better than other DLIRs investigated, both in terms of average DSC and variance of DSC. However, the improvement above other DLIR methods is more obvious where motions are greater (conditions where no-registration DSC is low, such as healthy hearts and LV hypertrophy) and vice versa (such as cardiac amyloidosis).

V. DISCUSSIONS AND CONCLUSION

Developing DLIR algorithms to extract clinical measurements from echo more accurately and precisely is an important endeavor. Precision issues are widely acknowledged [3], [4], and are considered a severe limiting factor in physicians' ability to improve care. Insufficient precision reduces our ability to diagnose and predict diseases accurately, and they

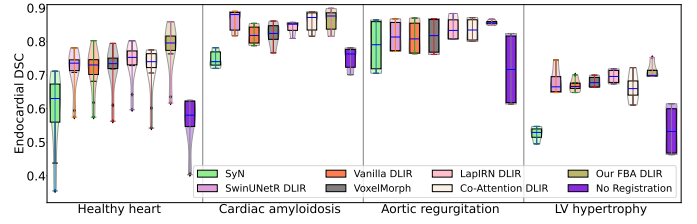


Fig. 8: Different DLIRs on the MITEA dataset to show how our FBA DLIR performs in the event of cardiac diseases.

further pose obstacles for benchmarking, comparison across different clinical centers, and improving guidelines.

Currently, the clinical norm is to use 2D echo, and although 3D/4D echo is now widely available via a 2D array transducer or via STIC imaging, they have yet to become mainstream approaches. This is likely because visualization and measurements in 3D are inconvenient, and there have yet to be significant trials to demonstrate definite advantages for diagnosis or prognosis. However, 2D imaging has disadvantages. For example, myocardial strains measured in 2D are biased due to the missing out-of-plane motion from 2D imaging [32], and in fetal echo, 2D imaging can have a misaligned imaging plane to cause normal hearts to be mistaken as defective and vice versa [33]. Further, current volume measurements in 2D in the clinic require very idealized assumptions of the cardiac chamber geometries and will have many inaccuracies [34]. Thus, it seems essential to develop DLIR for 3D echo. Recent work has shown that cardiac motions retrieved via 3D imaging can help predict patients' mortality risks, especially when processed with DL [35]. We thus maintain that 3D echo DLIR is the logical and essential next step in echo DLIR research.

Our experiments show that the FBA approach can be one way to enhance 3D DLIR performance. The key to FBA's ability to improve results is that it can evaluate areas of high registration errors and provide feedback to the DLIR to guide its improvement. A second registration is performed with this feedback, utilizing the co-attention architecture to guide this next DLIR in achieving the improvements. In the subsequent iteration, the network starts with the second registration from the previous iteration and uses its errors to provide feedback for further registration. In this sense, the multiple iterations during FBA training represent an iterative registration error convergence process (see Fig. 3), leading to improved registration results at the end of training (see Fig. 7).

Our results show that FBA works well with a range of existing 3D DLIR networks, providing non-overlapping benefits to existing networks, even those with attention mechanisms. It is thus a transferable strategy that can be added to promising DLIR networks. Our results further show that FBA DLIR works robustly under noise and naturally produces smooth surfaces on top of providing accurate results and that they can benefit both fetal and adult echo. We thus believe that it is essential to include the FBA approach when designing the state-of-the-art 3D DLIR.

FBA is consistent with the idea that there should be greater regularization of the warped moving image's quality for better

performance. In DLIR, although the resulting deformation field enables a high similarity metric between the fixed and warped images, the warped image often looks unnatural. We have previously shown that enhancing the warped images' quality, such as anatomic plausibility and image quality, can improve 2D echo DLIR [8]. Here, FBA represents a further regularization of the warped image to push it to have the right image features, matching those found in the fixed image.

Our results also suggest that 3D DLIR that utilizes attention generally does well, and attention may be enabling the development of the next-generation DLIR. This may be associated with challenges in scaling up from 2D to 3D. During this scaling up, having an additional dimension will naturally mean that the percentage of the image that contains the anatomical structures of interest becomes lower, and it will also mean that the number of image features that need to be considered will scale up significantly. Consequently, DLIR training will be more difficult and less effective, as much of the DLIR's effort will be expended on regions and features that are unimportant. The attention can address this sparse region and features of interest problem by guiding the DLIR training on important locations and features, helping it regain effectiveness.

REFERENCES

- [1] M. Azzam *et al.*, "3d-based strain analysis and cardiotoxicity detection in cancer patients received chemotherapy," *BMC cancer*, vol. 23, no. 1, p. 760, 2023.
- [2] S. P. Murphy *et al.*, "Heart failure with reduced ejection fraction: a review," *Jama*, vol. 324, no. 5, pp. 488–504, 2020.
- [3] E. Gayat *et al.*, "Reproducibility and inter-vendor variability of left ventricular deformation measurements by three-dimensional speckle-tracking echocardiography," *Journal of the American Society of Echocardiography*, vol. 24, no. 8, pp. 878–885, 2011.
- [4] N. H. van Oostrum *et al.*, "Fetal strain and strain rate during pregnancy measured with speckle tracking echocardiography: A systematic review," *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 250, pp. 178–187, 2020.
- [5] G. Balakrishnan *et al.*, "Voxelmorph: a learning framework for deformable medical image registration," *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.
- [6] S. S. Ahn *et al.*, "Co-attention spatial transformer network for unsupervised motion tracking and cardiac strain analysis in 3d echocardiography," *Medical Image Analysis*, vol. 84, p. 102711, 2023.
- [7] J. Hung *et al.*, "3d echocardiography: a review of the current status and future directions," *Journal of the American Society of Echocardiography*, vol. 20, no. 3, pp. 213–233, 2007.
- [8] M. K. Hasan *et al.*, "Multi-scale, data-driven and anatomically constrained deep learning image registration for adult and fetal echocardiography," *arXiv:2309.00831*, 2023.
- [9] C. Lian *et al.*, "Cocyclereg: Collaborative cycle-consistency method for multi-modal medical image registration," *Neurocomputing*, vol. 500, pp. 799–808, 2022.
- [10] S. Woo *et al.*, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
- [11] M.-M. Rohé *et al.*, "Svf-net: learning deformable image registration using shape matching," in *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I* 20. Springer, 2017, pp. 266–274.
- [12] X. Yang *et al.*, "Quicksilver: Fast predictive image registration—a deep learning approach," *NeuroImage*, vol. 158, pp. 378–396, 2017.
- [13] T. C. Mok and A. C. Chung, "Large deformation diffeomorphic image registration with laplacian pyramid networks," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III* 23. Springer, 2020, pp. 211–221.
- [14] R. R. Upendra *et al.*, "A convolutional neural network-based deformable image registration method for cardiac motion estimation from cine cardiac mr images," in *2020 Computing in Cardiology*. IEEE, 2020, pp. 1–4.
- [15] A. Østvik *et al.*, "Myocardial function imaging in echocardiography using deep learning," *IEEE transactions on medical imaging*, vol. 40, no. 5, pp. 1340–1351, 2021.
- [16] H. Wei *et al.*, "Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II* 23. Springer, 2020, pp. 623–632.
- [17] Z. Wang *et al.*, "Unsupervised echocardiography registration through patch-based mlps and transformers," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2022, pp. 168–178.
- [18] A. Lu *et al.*, "Learning-based regularization for cardiac strain analysis via domain adaptation," *IEEE transactions on medical imaging*, vol. 40, no. 9, pp. 2233–2245, 2021.
- [19] X. Lu *et al.*, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3623–3632.
- [20] H. Wang *et al.*, "A temporal attention based appearance model for video object segmentation," *Applied Intelligence*, pp. 1–11, 2022.
- [21] S. S. Ahn *et al.*, "Multi-frame attention network for left ventricle segmentation in 3d echocardiography," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I* 24. Springer, 2021, pp. 348–357.
- [22] L.-C. Chen *et al.*, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
- [23] J. Zbontar *et al.*, "Barlow twins: Self-supervised learning via redundancy reduction," in *International conference on machine learning*. PMLR, 2021, pp. 12 310–12 320.
- [24] T. Chen *et al.*, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
- [25] D. Zhao *et al.*, "Mitea: A dataset for machine learning segmentation of the left ventricle in 3d echocardiography using subject-specific labels from cardiac magnetic resonance imaging," *Frontiers in Cardiovascular Medicine*, vol. 9, p. 1016703, 2023.
- [26] H. Wiputra *et al.*, "Cardiac motion estimation from medical images: a regularisation framework applied on pairwise image registration displacement fields," *Scientific reports*, vol. 10, no. 1, p. 18510, 2020.
- [27] B. B. Avants *et al.*, "Advanced normalization tools (ants)," *Insight j*, vol. 2, no. 365, pp. 1–35, 2009.
- [28] S. Klein *et al.*, "Elastix: a toolbox for intensity-based medical image registration," *IEEE transactions on medical imaging*, vol. 29, no. 1, pp. 196–205, 2009.
- [29] A. Hatamizadeh *et al.*, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI Brainlesion Workshop*. Springer, 2021, pp. 272–284.
- [30] A. Hatamizadeh *et al.*, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.
- [31] J. Ternacle *et al.*, "Causes and consequences of longitudinal lv dysfunction assessed by 2d strain echocardiography in cardiac amyloidosis," *JACC: Cardiovascular Imaging*, vol. 9, no. 2, pp. 126–138, 2016.
- [32] M. Ren *et al.*, "Contribution of ventricular motion and sampling location to discrepancies in 2d versus 3d fetal ventricular strain measures," *Journal of the American Society of Echocardiography: Official Publication of the American Society of Echocardiography*, pp. S0894–7317, 2023.
- [33] L. Yeo, S. Luewan, and R. Romero, "Fetal intelligent navigation echocardiography (fine) detects 98% of congenital heart disease," *Journal of ultrasound in medicine*, vol. 37, no. 11, pp. 2577–2593, 2018.
- [34] V. C.-C. Wu and M. Takeuchi, "Three-dimensional echocardiography: current status and real-life applications," *Acta Cardiologica Sinica*, vol. 33, no. 2, p. 107, 2017.
- [35] G. A. Bello *et al.*, "Deep-learning cardiac motion analysis for human survival prediction," *Nature machine intelligence*, vol. 1, no. 2, pp. 95–104, 2019.