Feedback Attention for Unsupervised Cardiac Motion Estimation in 3D Fetal Echocardiography

Md. Kamrul Hasan, Guang Yang, and Choon Hwai Yap

Department of Bioengineering, Imperial College London, UK
{k.hasan22,g.yang,c.yap}@imperial.ac.uk

Abstract. Echocardiography motion estimation is essential for clinical cardiac health assessments, such as myocardial strain and ejection fraction. Attaining reliable performance with deep learning image registration (DLIR) is traditionally challenging due to intrinsic noise and fuzzy anatomic boundaries, but it is especially so for fetal echocardiography, where structures are smaller. It is further advantageous to achieve DLIR in 3D, as the fetal cardiac anatomy has complex 3D structures and motions that are difficult to understand with 2D platforms, especially with malformations. However, successful regularization strategies for 2D DLIR are often not as effective in 3D, and to date, 3D DLIR for fetal echo has not been implemented. Here, we propose a self-supervision module as regularization for the unsupervised DLIR network for 3D+time fetal echocardiography. We propose a novel feedback spatial transformer module where the registration outputs are used to generate a co-attention map that describes the remaining registration errors to guide the network's spatial emphasis during DLIR. This effectively facilitates self-supervision. This feedback attention approach improves results when added to existing transformer-based approaches, including the co-attention spatial transformer with and without the spatial and channel attention in the DLIR backbone, suggesting non-overlapping benefits with existing approaches. Further, our 3D results are consistent with our 2D DLIR investigations, where we find that a focus on the resulting image after warping with registration output is key to good performance. The code will be publicly available on GitHub (link will be provided after anonymous reviews (upon acceptance)).

Keywords: Echocardiography registration · Cardiac motion estimation · 4D (3D+time) fetal echocardiography · Feedback attention · Self-supervised image registration.

1 Introduction

The cardiac motion estimation from echocardiography is crucial for assessing cardiac function and identifying dysfunctions. Motion estimation can achieve comparable precision as speckle tracking in deriving myocardial (MYO) strains [1], which are important indicators of cardiac contractility health. In fetal echo, this is important for determining if the heart is sufficiently diseased to warrant interventions, such as placenta ablation for twin-to-twin transfusion syndrome [2] or catheter-based fetal heart intervention [3]. To date, fetal echo measurements have limited precision, which is poorer than postnatal and adult echo due to the smaller size of the fetal heart and the greater distance between the fetal heart and the transducer. For example, fetal MYO longitudinal strains for the right ventricle can vary by 2.5 times across different studies [4]. It is thus essential to develop effective deep learning (DL) approaches for resolving precision problems in fetal echo.

Further, although 3D+time fetal echo imaging is widely available, the clinical norm has remained with 2D echo due to the limited effectiveness of current 3D fetal echo visualization, where anatomic details are missed even by experts [5]. However, 2D imaging can have a misaligned imaging plane, causing normal hearts to be mistaken as defective and vice versa [6]. It is further difficult to mentally understand the complex 3D fetal cardiac anatomy and motion in 2D, especially with complex diseases with malformations, and 3D images can achieve more accurate volumetric measurements. To date, however, there has not been any DLIR algorithm applied to 3D fetal echo, which is our focus here.

Various DLIR techniques have been proposed and shown to have robust results for 2D echo. This includes the PWC-Net, which uses a feature pyramid extractor and an optical flow estimator [7] and a patch-based MLP and transformer network for 2D adult echo [8]. We have previously proposed a 2D DLIR incorporating an anatomic shape-encoder and an adversarial image texture constraint to a multi-scale DLIR [9]. However, techniques successful in 2D are often less effective in 3D due to the increased dimensionality and information in 3D images and the difficulty in scaling up training to 3D. To date, only one recent study demonstrated successful DLIR in 3D adult echo [10], where they proposed a co-attention spatial transformer to provide this necessary regularization. However, there remains room for improvement in the outcomes, and there are further challenges in fetal echo with regard to resolution and image quality.

Here, we propose to enhance existing DLIR transformer approaches by a novel feedback attention (FBA) module. We use the warped image obtained from the registration output to compute a co-attention map that indicates locations where registration errors remain high and insert this co-attention map as feedback to the registration network to guide performance enhancement. This is a logical next step and in line with our previous 2D DLIR studies, where we have found that an additional focus on the warped image quality is capable of improving existing DLIR strategies and is sufficient in achieving robust results. We show that FBA can similarly enhance 3D echo DLIR and can produce robust results in 3D fetal echo.

Attention-based approaches are gaining importance in medical image analysis [11]. Woo et al. [12] proposed a spatial and channel attention (SCA) approach to DL models to guide networks on what locations and features to emphasize, while Schlemper et al. [13] employed spatial attention in a gated module at each UNet [14] level for spatial attention effects. Co-attention can also be multi-frame, such as in [15]. Here, we show that our FBA approach can provide non-overlapping benefits with existing transformers and can thus be useful add-ons for synergistic effects. Our specific contributions are:

- We propose a new feedback attention (FBA) module to enhance spatial transformers for unsupervised 3D fetal echo DLIR, where the DLIR's outputs are used to compute a spatial co-attention map for feedback to the DLIR for additional self-supervision. We show that this improves 3D DLIR performance and show that this is due to an improved ability to correctly emphasize regions of interest.
- We couple our novel FBA to an SCA-based backbone network to extend our attention mechanism to be both location- and feature-based. We show that SCA has a non-overlapping advantage with FBA, and this coupling improves performance.

- We demonstrate that improved performance for 3D DLIR enables robust results in 3D fetal echo despite poorer resolution and image quality than an adult echo.

2 Methodology and Materials

2.1 Methods

Fig. 1 depicts our proposed DLIR pipeline, consisting of a SCA-based DLIR block (A) that predicts a deformation field from an image pair and two spatial attention blocks (B and C) that modify the input images to block A with a co-attention map between the pair. Block B generates the co-attention map between the original fixed and moving image to be applied to both images for the first registration block (A1) to highlight the spatial locale that needs emphasis during registration. Block C generates the co-attention map between the warped image (moving image deformed by initial registration output) and the fixed image, which is to be applied to the fixed image as feedback from the results for inputs into a second registration block (A2), providing self-supervision. This co-attention map in block C plays a crucial role in highlighting the spatial locales of dissimilarities between the warped and fixed images and, thus, of registration errors. Since block C uses the outputs of registration to modify its inputs, it serves as a form of feedback mechanism, making the model adaptable and responsive. Blocks A1 and A2 are identical networks with the same weights. More details are provided below.

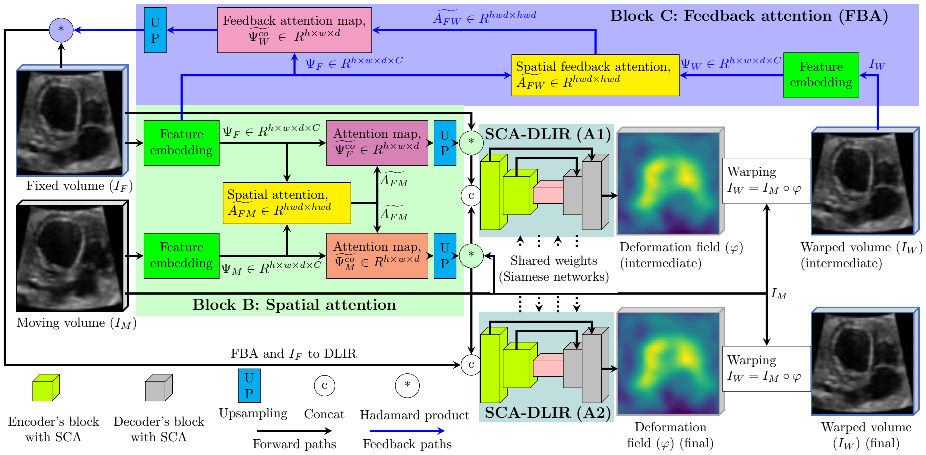
Block C: Feedback attention (FBA)
U P
Feedback attention map, $\widetilde{\Psi^{co}_W} \in R^{h\times w\times d}$
$\widetilde{A_{FW}} \in R^{hwd\times hwd}$
$\Psi_F \in R^{h\times w\times d\times C}$
Spatial feedback attention, $\widetilde{A_{FW}} \in R^{hwd\times hwd}$
$\Psi_W \in R^{h\times w\times d\times C}$
Feature embedding
I_W
*
Fixed volume (I_F)
Feature embedding
$\Psi_F \in R^{h\times w\times d\times C}$
Attention map, $\widetilde{\Psi^{co}_F} \in R^{h\times w\times d}$
U P
Spatial attention, $\widetilde{A_{FM}} \in R^{hwd\times hwd}$
$\widetilde{A_{FM}}$
$\widetilde{A_{FM}}$
Feature embedding
$\Psi_M \in R^{h\times w\times d\times C}$
Attention map, $\widetilde{\Psi^{co}_M} \in R^{h\times w\times d}$
U P
Block B: Spatial attention
Moving volume (I_M)
SCA-DLIR (A1)
Shared weights (Siamese networks)
Deformation field (φ) (intermediate)
Warping $I_W = I_M \circ \varphi$
Warped volume (I_W) (intermediate)
I_M
FBA and I_F to DLIR
c
SCA-DLIR (A2)
Deformation field (φ) (final)
Warping $I_W = I_M \circ \varphi$
Warped volume (I_W) (final)
Encoder's block with SCA
Decoder's block with SCA
Upsampling
Concat
Hadamard product
Forward paths
Feedback paths

Fig. 1: The proposed self-supervised DLIR pipeline using our novel FBA (block C) module, incorporating Siamese SCA-DLIR networks (block A1 and A2) and spatial attention (block B). Our FBA module facilitates the SCA-DLIR's self-supervision.

The unsupervised DLIR networks. The backbone DLIR (Block A1 and A2 in Fig. 1) consists of three encoding and decoding blocks in the UNet structure ($\mathcal{F}_{\mathcal{CNN}}$), with

128, 256, and 512 filters. $\mathcal{F}_{\mathcal{CNN}}$ predict a deformation field $u(x)$ from a pair of N-dimensional input images (DDF, φ), $u(x) = \mathcal{F}_{\mathcal{CNN}}(I_F, I_M; \Theta)$; where Θ is training parameters and $u(x) : \mathcal{R}^N \mapsto \mathcal{R}^N$. To highlight the crucial regions and focus on relevant features across the channels, we use SCA [12] in the encoder and decoder of DLIR (SCA-DLIR). Blocks A1 and A2 are Siamese blocks using the same weights, where A1 is used for the initial registration between the fixed and moving images, and A2 is used for the subsequent registration after feedback is applied to the fixed image.

To train $\mathcal{F}_{\mathcal{CNN}}$, $\widehat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \left\{ \mathcal{L}(I_F, I_M \circ \varphi) \right\}$ is optimized, where $\mathcal{L}$ is a similarity metric such as cross-entropy and mutual information. However, this DLIR approach does not produce optimal results, which could, in part, be due to fetal echo's challenges in resolution, image quality, and small cardiac structures. Therefore, we propose to use our novel FBA as additional self-supervision and spatial co-attention to improve it.

FBA for self-supervision and Spatial attention. We use images near the end-systolic (ES) as the fixed image ($I_F \in \mathbb{R}^{H \times W \times D}$) and images near the end-diastole (ED) as the moving image ($I_M \in \mathbb{R}^{H \times W \times D}$), aiming to find a DDF to warp the I_M onto the I_F as $I_W = I_M \circ \varphi \in \mathbb{R}^{H \times W \times D}$, where H, W, and D indicate height, width, and depth of the 3D volume, respectively. In blocks B and C, we compute the 3D co-attention map [16] via feature embedding to direct our DLIR to emphasize coherent anatomical locations. This enables the DLIR to understand the association between two images in a global semantic feature space. The embedding network ($\mathcal{E}$) consists of a series of 3D convolutions having filters of 64, 128, and 256 and an atrous spatial pyramid pooling [17] module to use parallel dilated convolutions with different rates (6, 12, and 18) to capture multi-scale contextual information. For a given input $I \in \mathbb{R}^{H \times W \times D}$, $\mathcal{E}$ returns an embedded features map of $\Psi = \mathcal{E}(I) \in \mathbb{R}^{C(=256) \times h \times w \times d}$, where $h, w, d = H/8, W/8, D/8$. Algorithm 1 explains the procedures of spatial FBA (ϑ) and spatial attention (ξ) estimation.

The obtained spatial co-attention maps (ξ_F and ξ_M) and the FBA co-attention map (ϑ_W) are then applied to the input images to highlight locations of importance. For the first registration in block A1, co-attention maps from block B are applied to the original images, I_F and I_M, as $I'_F = I_F \times \xi_F$, and $I'_M = I_M \times \xi_M \in \mathbb{R}^{H \times W \times D}$, while in the second registration in block A2, the co-attention map from block C is applied to the fixed image, as $I''_F = I_F \times \vartheta_W \in \mathbb{R}^{H \times W \times D}$, while the moving image retains its modification by the co-attention map from block B. The final loss function to train our DLIR is in Eq. 1, where ϑ_W provides self-supervision as feedback.

$$\begin{aligned} \widehat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \Big\{ & \mathcal{L}_1(I_F, I_M \circ \varphi) + \mathcal{L}_1(I'_F, I'_M \circ \varphi) + \mathcal{L}_1(I''_F, I'_M \circ \varphi) \\ & + \mathcal{L}_2(\xi_F, \xi_M \circ \varphi) + \mathcal{L}_2(\vartheta_W, \xi_M \circ \varphi) \Big\}, \end{aligned} \tag{1}$$

where $\mathcal{L}_1$ is mean square error and $\mathcal{L}_2$ is normalized cross correlation.

2.2 Dataset and training

Dataset. Our fetal dataset is a 4D (3D+time) multi-demographic fetal echo collection. Ethics approval was obtained under Kepler University Hospital IRB protocol 1009/2017,

Algorithm 1: Estimation of spatial FBA and spatial attention.

```
Input: An image pair $\{I_F, I_M\}$ or $\{I_F, I_W\} \in \mathbb{R}^{H \times W \times D}$
Output: Spatial attention ($\xi \in \mathbb{R}^{H \times W \times D}$) using $\{I_F, I_M\}$ or spatial FBA ($\vartheta \in \mathbb{R}^{H \times W \times D}$)
        using $\{I_F, I_W\}$
Extract encoded semantic feature maps $\Psi_F = \mathcal{E}(I_F)$, $\Psi_M = \mathcal{E}(I_M)$, and $\Psi_W = \mathcal{E}(I_W)$
  $\in \mathbb{R}^{C \times h \times w \times d}$, where $\mathcal{E}$ is a feature embedding network and $C$ is the channel numbers
Compute affinity (similarity) matrix ($A$) to mine the correlations between $\Psi_F$ and $\Psi_M$ or
  $\Psi_F$ and $\Psi_W$ in their feature embedding space by flattening them ($\in \mathbb{R}^{C \times hwd}$),
      $A_{FM} = \Psi_F^T \cdot W_{fm} \cdot \Psi_M \in \mathbb{R}^{hwd \times hwd}$,
      $A_{FW} = \Psi_F^T \cdot W_{fw} \cdot \Psi_W \in \mathbb{R}^{hwd \times hwd}$,
  where $W \in \mathbb{R}^{C \times C}$ is a diagonal weight matrix.
Normalize the affinity matrix using a softmax activation ($\sigma_1$) as $\widetilde{A_{FM}} = \sigma_1(A_{FM})$ or
  $\widetilde{A_{FW}} = \sigma_1(A_{FW})$, which reflects the relevance of each feature ($k \in \{1, 2, ..., hwd\}$) in
  $\Psi_F$ to the $k^{th}$ feature in $\Psi_M$ or $\Psi_W$
Estimate co-attention enhanced attention feature map as: $\Psi_F^{co} = \Psi_F \cdot \widetilde{A_{FM}} \in \mathbb{R}^{C \times hwd}$ (for
  fixed image) and $\Psi_M^{co} = \Psi_M \cdot \widetilde{A_{FM}} \in \mathbb{R}^{C \times hwd}$ (for moving image), and
  $\Psi_W^{co} = \Psi_F \cdot \widetilde{A_{FW}} \in \mathbb{R}^{C \times hwd}$ (for feedback attention).
Apply post-processing to those co-attention enhanced feature maps (unflattened
  $\in \mathbb{R}^{C \times h \times w \times d}$) using a convolutional kernel ($\gamma$), bias ($\Omega$), and sigmoid activation ($\sigma_2$)
  with an output channel of 1: $\widetilde{\Psi^{co}} = \sigma_2(\gamma \Psi^{co} + \Omega) \in [0, 1]$, where $\widetilde{\Psi^{co}} \in \mathbb{R}^{h \times w \times d}$
Upsampling the refined attention maps using a trilinear interpolation ($\Upsilon$) provides:
  $\xi_F = \Upsilon(\widetilde{\Psi_F^{co}}) \in \mathbb{R}^{H \times W \times D}$, $\xi_M = \Upsilon(\widetilde{\Psi_M^{co}}) \in \mathbb{R}^{H \times W \times D}$, and $\vartheta_W = \Upsilon(\widetilde{\Psi_W^{co}}) \in \mathbb{R}^{H \times W \times D}$
```

and the National University of Singapore DSRB protocol 2014/00056, and written informed consent was obtained from all patients. Images are acquired with the GE Volusion 730 ultrasound machine under the STIC mode with the RAB $4-8L$ transducer and have an in-plane image resolution of $0.95\mu m \times 0.90\mu m$ (see details in [18]), as well as 25 to 40-time frames in each 4D volume image file. The database comprises 26 patients, consisting of healthy fetal patients and fetal patients with fetal aortic stenosis disease. The epicardial (EPI) and endocardial (ENDO) boundaries and the MYO tissue space are manually labeled at the ES and ED time points and carefully smoothed with as little compromise to segmentation accuracy as possible, and a validated cardiac motion estimation algorithm (multi-scale Elastix with cyclic regularization) is used to propagate the labels to other time points [18, 19]. A total of 78 fixed and moving pairs were generated by various combinations of the time frames near the ES and those near the ED. 69 (23 patients) are used for training and validation, and 9 (3 patients) are used for testing.

Training and evaluation. A batch of 4 and epochs of 200 are used with an Adam optimizer with a learning rate of 1e-5. All the volumetric 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 results are evaluated using a dice similarity coefficient (DSC) and mean square error (MSE) between the fixed and warped volumes.

3 Results and Discussions

Different DLIR configurations are used for an ablation study: **1. Van-DLIR** (Vanilla DLIR with only fixed and warped image similarity constraint), **2. Van-VoxelMorph** (Van-DLIR with myocardium DSC loss constraint [20]), **3. SCA-DLIR** (Van-DLIR with SCA [12]), **4. SCA-VoxelMorph** (SCA-DLIR with myocardium DSC loss constraint), **5. CoVan-DLIR** (Van-DLIR with spatial co-attention [10]), **6. FBA-Van-DLIR** (Van-DLIR with FBA), and **7. FBA-SCA-DLIR** (SCA-DLIR with FBA).

The results in Table 1 show that adding SCA to the Van-DLIR improves ($p \ll 0.05$) the anatomical benchmarks (DSC) while having identical texture measures (MSE), indicating that the SCA mildly but significantly improves registration accuracy. However, adding myocardial DSC-based auxiliary loss to Van-DLIR or SCA-DLIR does not improve DSC. Our previous experience is that DSC-loss can improve Van-DLIR in 2D echo [9]; as such, this lack of improvement is likely due to a 3D image being the subject here and can demonstrate how successful strategies in 2D DLIR are often not effective in 3D. When we add spatial co-attention to Van-DLIR (CoVan-DLIR), DSC improves for MYO ($p \ll 0.05$) and EPI from Van-DLIR but not for ENDO. With CoVan-DLIR, DSC results are generally comparable to those of SCA-DLIR. However, when our novel FBA is added to Van-DLIR, it provides improvements beyond SCA, co-attention, and VoxelMorph, where DSC is significantly improved for all three regions ($p \ll 0.05$). CoVAN-DLIR has recently been proposed in [10] as the first unsupervised 3D DLIR, demonstrated using adult echo, and is thus arguably the state-of-the-art. It enables the emphasis of specific spatial regions during DLIR to enable an enhancement from Van-DLIR. Our results show that when applied to 3D fetal echo, it provides performance similar to SCA, suggesting that SCA can similarly tap into the spatial attention effect.

Table 1: The quantitative registration results to show the non-overlapping benefits of each integrated module. Bold fonts indicate the best metrics.

Registration methods	**DSC**			**MSE**
	MYO	**ENDO**	**EPI**	
Van-DLIR	0.832 ± 0.031	0.907 ± 0.009	0.870 ± 0.020	**0.004 ± 0.001**
Van-VoxelMorph [20]	0.839 ± 0.029	0.899 ± 0.013	0.869 ± 0.021	0.005 ± 0.002
SCA-DLIR	0.840 ± 0.016	0.913 ± 0.008	0.877 ± 0.012	**0.004 ± 0.001**
SCA-VoxelMorph	0.841 ± 0.009	0.898 ± 0.008	0.870 ± 0.008	0.005 ± 0.002
CoVan-DLIR [10]	0.845 ± 0.009	0.903 ± 0.005	0.874 ± 0.007	0.005 ± 0.001
FBA-Van-DLIR	0.850 ± 0.025	0.911 ± 0.013	0.881 ± 0.019	0.005 ± 0.002
FBA-SCA-DLIR	**0.864 ± 0.025**	**0.921 ± 0.015**	**0.893 ± 0.019**	0.005 ± 0.002

However, our results show that enhancing CoVan-DLIR with our novel FBA can improve CoVan-DLIR even further. FBA adds a feedback co-attention block to CoVan-DLIR and repeats the registration process after the feedback. In FBA, a new co-attention map is computed between the warped and fixed images to be applied to the fixed image for the second DLIR. The warped image is the transformation of the moving image towards the fixed image, and the DLIR is trained to match the warped image to the fixed image as much as possible. Our FBA co-attention map compares these two images that are supposed to match well; thus, it is a map to highlight locations with remaining registration matching errors. This thus gives FBA a new capability to more intelligently figure out where to focus further to minimize registration errors and can be the reason for its better performance. In this sense, FBA can also be seen as a form of self-supervision for DLIR, where it provides feedback to the DLIR to highlight where there are still errors and to help guide the network to successful registration. In each iteration of the training, FBA uses the outputs of a first registration (block A1 in Fig. 1) to compute a feedback co-attention map, which is used for a second registration (block A2 in Fig. 1) before losses are computed. 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 (Fig. 1 in the supplemental material). Our results further show that FBA can similarly improve on SCA-DLIR, similar to the improvement it can provide to CoVan-DLIR. This suggests a non-overlapping benefit with SCA. However, FBA-SCA-DLIR has better DSC results than FBA-Van-DLIR, suggesting that FBA is more compatible with SCA than CoVan-DLIR.

Results in Fig. 2 show the reconstructed LV volume in the warped image under the various DLIRs. It can be observed that FBA results in warped volumes with fewer ENDO/ EPI border irregularity and abnormal spikes, with a visually good fit with the fixed volume, confirming results in Table 1. Fig. 3 shows the co-attention map multiplied by the fixed image to visualize the attention adopted by the DLIRs. With CoVan-DLIR alone, the part of the MYO wall is blurred (and yellow, denoting a lack of emphasis), such as at the top of the images within the red box, which bounds the LV). This shows that CoVan-DLIR's attention map did not fully retrieve the MYO topology. In contrast, FBA's attention map recovers a well-visible MYO structure (in blue, which presents strong emphasis). This can enable better learning of the registration feature and, consequently, better DLIR. Fig. 4 shows learned feature maps from Van-DLIR and our FBA-SCA-DLIR. Here, we can observe that Van-DLIR feature maps do not preserve the MYO topology, but FBA-SCA-DLIR feature maps clearly display the structure of the MYO, showing that our proposed attention approach is successful at bringing emphasis to the relevant structures and that this is likely the basis for the observed improved results.

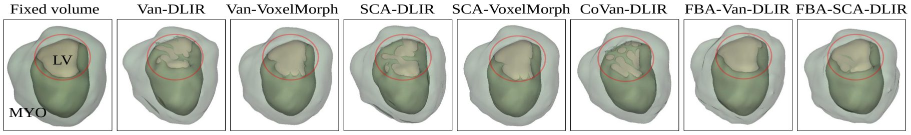

Fig. 2: Warped volumes from different DLIRs for the same testing patient, demonstrating ENDO/EPI border irregularity and abnormal spikes (inside the red circle).

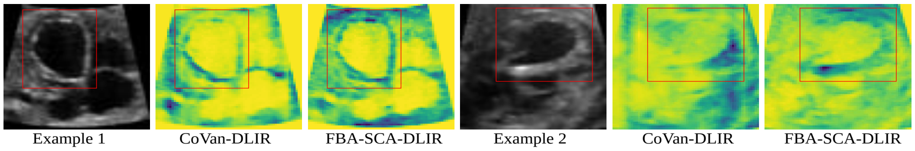

Fig. 3: Visualization of attention maps as heatmaps overlaid on the raw images, showing the DLIR's attention focus (Yellow: less focus and Blue: high focus). The MYO structures are well visible (see red box) in our FBA-SCA-DLIR compared to CoVan-DLIR.

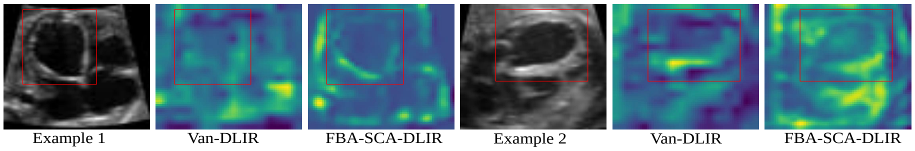

Fig. 4: Visualization of feature maps with corresponding raw images from identical DLIR layers for Van-DLIR and our proposed FBA-SCA-DLIR, showing that our DLIR retrieves MYO's structural information in the feature maps (see red box).

Our study thus confirms that this ability to place specific spatial and feature emphasis to guide DLIR is essential in 3D. The additional dimension in 3D images compared to 2D images means that the percentage of the image that contains structures of interest becomes lower, and DLIR training will consequently be more difficult and less effective. The attention mechanism that guides DLIR training towards important locations and features can overcome this sparser region of interest problem to offer better performance.

4 Conclusion

We proposed a novel FBA module to provide self-supervision to the 3D DLIR and showed that this improves DLIR. FBA enhances the guidance provided to DLIR by emphasizing locations where registration errors remain high to help the network improve accuracy and is equivalent to an iterative error convergence during training. When FBA is combined with an SCA-DLIR backbone to enable both spatial and feature attention, results are further improved, and robust DLIR results are observed for fetal echo despite challenges in resolution, image quality, and small size of cardiac structures. Our results further show that co-attention is an ideal approach to scaling up from 2D DLIR to 3D, which can be challenging for several successful 2D DLIR strategies. Our proposed network can thus contribute to advancing state-of-the-art fetal echo motion tracking, potentially leading to a more accurate and reliable assessment of fetal cardiac function.

Acknowledgements. M. K. Hasan was supported by the EPSRC-DTP studentship funds from the 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, the SABER project supported by Boehringer Ingelheim Ltd, and the UKRI Future Leaders Fellowship (MR/V023799/1).

References

[1] B. Heyde, R. Jasaityte, D. Barbosa, V. Robesyn, S. Bouchez, P. Wouters, F. Maes, P. Claus, J. D'hooge, Elastic image registration versus speckle tracking for 2-d myocardial motion estimation: a direct comparison in vivo, IEEE transactions on medical imaging 32 (2012) 449–459.

[2] R. Cincotta, S. Kumar, Future directions in the management of twin-to-twin transfusion syndrome, Twin Research and Human Genetics 19 (2016) 285–291.

[3] A. Tulzer, W. Arzt, R. Gitter, E. Sames-Dolzer, M. Kreuzer, R. Mair, G. Tulzer, Valvuloplasty in 103 fetuses with critical aortic stenosis: outcome and new predictors for postnatal circulation, Ultrasound in Obstetrics & Gynecology 59 (2022) 633–641.

[4] N. H. van Oostrum, C. M. de Vet, D. A. van der Woude, H. M. Kemps, S. G. Oei, J. O. van Laar, Fetal strain and strain rate during pregnancy measured with speckle tracking echocardiography: A systematic review, European Journal of Obstetrics & Gynecology and Reproductive Biology 250 (2020) 178–187.

[5] B. Adriaanse, C. Tromp, J. Simpson, T. Van Mieghem, W. Kist, D. Kuik, D. Oepkes, J. Van Vugt, M. Haak, Interobserver agreement in detailed prenatal diagnosis of congenital heart disease by telemedicine using four-dimensional ultrasound with spatiotemporal image correlation, Ultrasound in obstetrics & gynecology 39 (2012) 203–209.

[6] L. Yeo, S. Luewan, R. Romero, Fetal intelligent navigation echocardiography (fine) detects 98% of congenital heart disease, Journal of ultrasound in medicine 37 (2018) 2577–2593.

[7] A. Østvik, I. M. Salte, E. Smistad, T. M. Nguyen, D. Melichova, H. Brunvand, K. Haugaa, T. Edvardsen, B. Grenne, L. Lovstakken, Myocardial function imaging in echocardiography using deep learning, ieee transactions on medical imaging 40 (2021) 1340–1351.

[8] Z. Wang, Y. Yang, M. Sermesant, H. Delingette, Unsupervised echocardiography registration through patch-based mlps and transformers, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer, pp. 168–178.

[9] M. K. Hasan, H. Zhu, G. Yang, C. H. Yap, Multi-scale, data-driven and anatomically constrained deep learning image registration for adult and fetal echocardiography, arXiv preprint arXiv:2309.00831 (2023).

[10] S. S. Ahn, K. Ta, S. L. Thorn, J. A. Onofrey, I. H. Melvinsdottir, S. Lee, J. Langdon, A. J. Sinusas, J. S. Duncan, Co-attention spatial transformer network for unsupervised motion tracking and cardiac strain analysis in 3d echocardiography, Medical Image Analysis 84 (2023) 102711.

[11] X. Li, M. Li, P. Yan, G. Li, Y. Jiang, H. Luo, S. Yin, Deep learning attention mechanism in medical image analysis: Basics and beyonds, International Journal of Network Dynamics and Intelligence (2023) 93–116.

[12] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), pp. 3–19.

[13] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, D. Rueckert, Attention gated networks: Learning to leverage salient regions in medical images, Medical image analysis 53 (2019) 197–207.

[14] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, pp. 234–241.

[15] S. S. Ahn, K. Ta, S. Thorn, J. Langdon, A. J. Sinusas, J. S. Duncan, 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, pp. 348–357.

[16] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, F. Porikli, 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, pp. 3623–3632.

[17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE transactions on pattern analysis and machine intelligence 40 (2017) 834–848.

[18] H. Wiputra, W. X. Chan, Y. Y. Foo, S. Ho, C. H. Yap, Cardiac motion estimation from medical images: a regularisation framework applied on pairwise image registration displacement fields, Scientific reports 10 (2020) 18510.

[19] W. X. Chan, Y. Zheng, H. Wiputra, H. L. Leo, C. H. Yap, Full cardiac cycle asynchronous temporal compounding of 3d echocardiography images, Medical Image Analysis 74 (2021) 102229.

[20] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, A. V. Dalca, Voxelmorph: a learning framework for deformable medical image registration, IEEE transactions on medical imaging 38 (2019) 1788–1800.