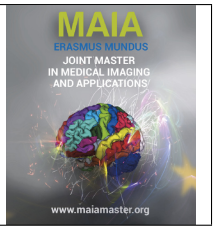
Medical Imaging and Applications

Master Thesis, June 2023



Temporal Image Registration using deep learning for 3D Fetal Echocardiography

Kazi Saeed Alam, Md Kamrul Hasan, Dr Choon Hwai Yap

Department of Bioengineering, Imperial College London, UK

Abstract

The fetal heart can experience congenital heart malformation and functional abnormalities. Ultrasound imaging plays a vital role in assessing the heart structure and function of the developing fetus due to its non-incisive nature. However, the detection of such abnormalities via mass screening is only 50%, suggesting a need for further improvement. Many researchers have been working in order to detect abnormalities in the heart from ultrasound imaging through segmenting cardiac chambers, valves, and blood flow patterns but most of the works are based on adult hearts. This motivates us to explore fetal echocardiographic images for which we collected 4D volume fetal heart images to perform temporal registration to segment the myocardium and left ventricle chamber from these images. Having a deep learning-enabled standardized approach to evaluation can improve precision and accuracy. Thus, in this project, we propose to develop methods for automatic 3D segmentation based on temporal registration from 4D fetal echo images. The 4D fetal echo images were collected and properly annotated with the help of an existing cardiac motion estimation algorithm. Our proposed model is built upon UNET based image registration model as a baseline with the residual branch, which is guided by a variational autoencoder to enforce structural features of the heart via latent space training and adversarial learning. We also plan to make the proposed model perform multi-scale registration. We have developed and tested our proposed network for both 2D (Adult images from CAMUS Dataset) and 3D (Fetal Data) segmentation which showed significant performance in both cases. As evaluation metrics, Mean squared error, and Dice Metric were computed both before and after the registration process.

Keywords: Fetal Echocardiography, Ultrasound, Image Registration, Variational Autoencoder, Adversarial Learning

1. Introduction

Ultrasound is one of the major imaging techniques that play a vital role to monitor cardiac functions and abnormalities. Due to its non-invasive nature ultrasound imaging has gained much popularity and has been used to assess heart structure and function by monitoring cardiac chambers, valves, and blood flow patterns. This imaging modality enables clinicians to diagnose and monitor congenital heart defects, providing valuable information for early intervention and management. Ciancarella et al. (2020), Sachdeva and Gupta (2020) showed the significance of the use of ultrasound in the field of cardiac imaging.

Heart structure and shapes such as Cardiac chambers, valves, blood flow patterns, etc can be used as good identifiers to detect and evaluate several cardiac diseases like Congenital heart defects, Coronary artery disease, Valvular heart diseases, Cardiomyopathies, etc. Research works from Green et al. (2023), Ong et al. (2020) shows that the information gained from the shape of the myocardium and heart chambers can give valuable insight which can detect and evaluate various Congenital heart defects. Researchers have been working to improve the detection of these diseases by employing automatic detection of heart structures and shapes. Having a deep learning-enabled standardized approach to automatically segment and detect can improve precision and accuracy.

Although most of the works are based on assessing the adult heart, the fetal heart also can experience congenital heart malformation and functional abnormalities. However, the detection of such problems via mass screening is only around 50%, suggesting a need for improvement. Being surprised at birth with fetal heart abnormalities instead of detecting them during mid-

gestation reduces the time available for planning and executing surgical treatment, and leads to poorer outcomes. Further, the evaluation approach of evaluating fetal heart health via fetal echo depends on many manual processes and involves subjective interpretation.

In our work, we are proposing a 3D temporal image registration-based segmentation technique to automatically detect and assess the left ventricle heart chamber and myocardium. The novelty of this research is mainly:

- A whole new 4D fetal echocardiography dataset with annotated 3D LV and myocardium masks for each 3D volume image. There is less research on fetal heart echocardiography due to the scarcity of well-produced datasets. We proposed an efficient workflow to manually segment the heart chamber and myocardium with temporal registration. An existing cardiac motion estimation algorithm has been used to assist the algorithm development. We hope that the publication of this new dataset will create a benchmark for further fetal heart echocardiography analysis and assessment.

- As the estimation of the deformation field by registration between two time points can help share the information between two segmentation branches, we are proposing a robust and efficient technique for temporal image registration for 4D fetal echocardiogram image volume.

We have proposed a Multi-class Anatomically Constrained and Multi-scale Registration (MACMR) framework in our research. The proposed registration method has the following integral parts:

- **Vanilla-DLIR**: The baseline architecture of the temporal registration is based on the typical UNET-like structure for performing image segmentation. The use of residual blocks helps to avoid the degradation of the features' quality as a non-zero regularizing path will skip over them. We are calling this baseline model Vanilla DLIR (Deep learning based image registration) as here the encoders of UNET try to extract features from lower to higher space and pass to the bottleneck whereas the task of the decoder is to produce the deformation field for the moving image so that it can be warped to match as much as possible as the target image.

- **AC-DLIR**: We have proposed to include a Variational encoder to enforce structural features of the heart via latent space training. The local segmentation-aware loss (fixed and moved labels) uses pixel-level predictions and may not ensure a satisfactory global match between the warped source and target anatomical masks. For this reason, the global latent space features can be beneficial for the network to perform better.

- **AdvAC-DLIR**: Moreover, we also propose to include adversarial learning as like zero-sum game theory (one agent's gain is another agent's loss), where the discriminator is used to classify moved and fixed images.

Still, there is room for performance improvement. Hence, we proposed Multi-class Anatomically Constrained and Multi-scale Registration (MACMR) framework. Additionally, we need a registration framework that can provide a suitable deformation field for all the scales of decoders in the proposed segmentation network to share the motion information. We have evaluated our proposed model for 2D as well as 3D volume datasets. For 2D data, the CAMUS 2D adult Echocardiography data were used from Leclerc et al. (2019) whereas for 3D data, the proposed fetal dataset. In order to validate our proposed framework, we have conducted several experiments on the existing DL-based registration pipeline.

2. Literature Review

Researchers have worked in the field of medical image registration in various directions. A broad topic like image registration can be classified into various objectives. The methods can be interpatient or intra-patient (same patient at different time points), and the images can be from one single imaging technique (unimodal) or can be of multimodal techniques. The registration methods can be deformable, affine, or simply rigid. Also based on the organ of interest, it can be the brain, lungs, heart, or even tumors and so on. Input images can be of different types of dimensions or combinations of them. In our work, we have tried to cover the unimodal, intrapatient fetal echocardiographic registration based on 4D volume images. In the following sections, the recent trends in image registration as well as focus based on ultrasound techniques will be explored.

2.1. Deep learning based Image Registration

We will restrict the discussion of trends in medical image registration in deep learning-based (DL) techniques as the recent works have shown an upward trend in the domain of image registration yielding state-of-the-art for various applications. The use of conventional similarity-based metrics such as mean-squared error, structural similarity, cross-correlation, mutual information, etc work well for unimodal image registration in the case of CT or MRI images, as shown in Gong et al. (2017), Heinrich et al. (2012). But the presence of noise such as in ultrasound images or in the case

of multi-modal registration they failed to perform satisfactorily (Rivaz et al. (2014)). Many researchers have replaced these conventional methods with CNN-based deep-learning image registration and achieved success.

Cheng et al. (2018) proposed an unsupervised learning-based registration method to train a classifier to learn the deformation field using continuous probabilistic values for similarity measures. In their work, they have claimed the learned deep similarity metric outperforms MI as in conventional methods in brain T1-T2 registration. The challenge was to have a smooth first-order derivative to have a better overlap between the fixed and moving images. Other works from Simonovsky et al. (2016), Ferrante et al. (2018) also explored the use of deep similarity metrics with unsupervised or weakly supervised training. The challenge of these works was to acquire an accurately aligned image. Images with noise such as ultrasound or in the case of multi-modal image pairs, the same performance will be difficult to achieve.

Compared to other modalities, ultrasound images are a bit challenging due to the image acquisition technique and also due to the presence of artifacts such as speckle noise. Haskins et al. (2018) in their work showed the comparison of multimodal image registration based on deep learning similarities. They have shown the CT-MRI pairs have better registration than the MRI-US pair in the case of the use of single similarity metrics. Ferrante et al. (2018) proposed the use of multiple metrics instead of single ones and showed improved performance for ultrasound image registration.

Wu et al. (2016) have introduced the use of variational autoencoders to perform latent space training, they have shown the use of both local and global features improved the performance of training with only local features. They have used the segmented masks as well as intensity images of brain MRI to perform the latent space training for image registration. They used a stack of autoencoders for the model to learn the latent space features and compared the result with Dice Similarity Coefficient (DSC). Although the dice similarity between the masks improves the smoothness of the shapes of the human organ, the intensity similarity between fixed and moving image still needs further improvement.

To provide better regularization which was lacking in the works discussed by VAs, some researchers proposed the use of adversarial learning or GAN-based models. As human organs are highly regular, better regularization is needed to have plausible shapes in the produced output. Yan et al. (2018) in their work have trained GAN-based networks to discriminate between ground truth-based and prediction-based transform to deform images. In their work, they have used the adversarial loss to optimize the accurate transform to deform the fixed image. Fu et al. (2020) also showed similar improvement in registration performance by introducing adversarial loss. GAN-based models helped to generate more plausible and medically acceptable structures and shapes after registration in these research works. However, the similarity in intensity matching for GANs still needs to be investigated thoroughly. Some recent advances also show the use of transfer learning, LSTMs, one-shot predictions, Faster RCNN, etc in Xie et al. (2022), Fechter and Baltas (2019), Jaderberg et al. (2015). However, all these have been applied to mostly CT and MRI images. While applying ultrasound images, most of them do not show any satisfactory improvement.

2.1.1. Image Registration in EchoCardiography

For cardiac chamber segmentation, ultrasound images can be acquired in two chambers (A2C) or four chambers (A4C) view. An optical flow estimation-based technique for deep, fully convolutional networks was suggested by Jafari et al. (2018). Jafari et al. (2019) also proposed the use of semi-supervised learning where they have incorporated inverse mapping with the use of adversarial learning and inverse mapping of the moved and target masks for LV segmentation. Yoon et al. (2021) in their work showed the use of Regional-CNN to extract geometrical attributes to perform LV segmentation. Variational autoencoders as discussed before have been also used for cardiac chamber segmentation tasks in cardiac ultrasounds. Painchaud et al. (2019) used VAs to represent the latent space training.

Some works have been also done in fetal echocardiography. Yang et al. (2020) in their work have used DeeplabV3 with UNET to segment the left ventricle chambers for the fetus's heart. Dong et al. (2019) also worked with A4C view for residual visual block network-based segmentation of the fetal heart chamber. But as the dataset is very limited for fetal echocardiography, still the performance of these models needs to be investigated further.

3. Dataset Description

4D echocardiography dicom images were acquired for studying out of which 4 were healthy fetuses and the rest were diseased cases. The fetuses were of mixed gender and different ethnic groups (Chinese, Indian, Malay). Most of the cases had a gestation age between 22 to 32 weeks. The images were obtained in accordance with protocol 2014/00056 from Domain Specific Review Board and with the consent of all the participants. The 4D echo images were carried out with GE Voluson 730 ultrasound connected to the RAB 4-8L transducer (GE Healthcare Inc., Chicago, Illinois, USA) which has approximated 154 μm axial resolution and around 219 μm lateral resolution along with a transducer of 5 MHz.

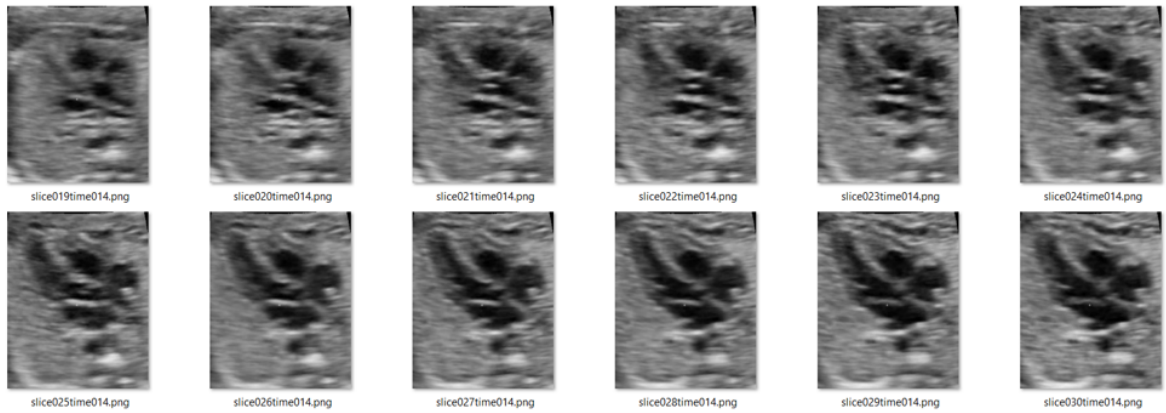


Figure 1: Visualization of intensity image slices (Patient001).

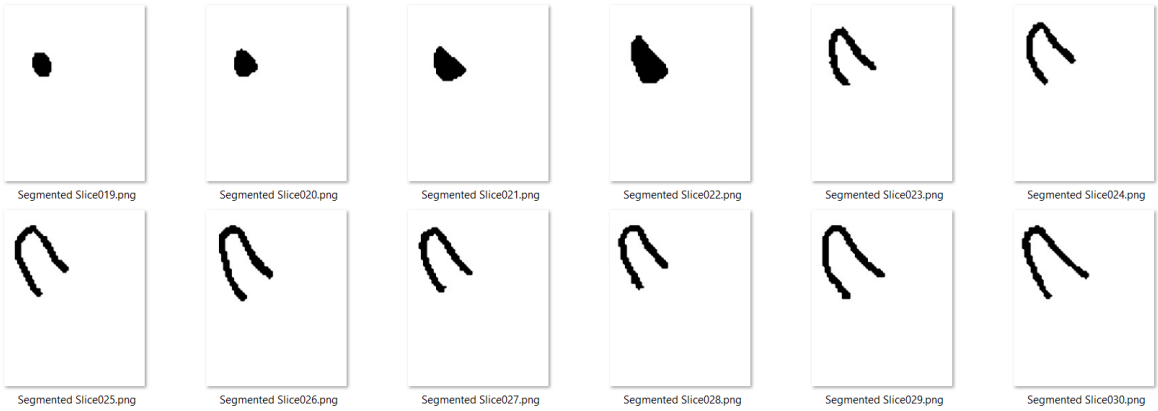


Figure 2: Visualization of annotated masks (Patient001).

3.1. Data Preparation

As the ultrasound raw images acquired were in 4D dicom format, they needed to be transformed to 3D format. This will help to extract the time points for each patient's case. To do so, a special software named "*4D View*" developed by GE Healthcare was used. This step will generate a cine sequence or cine loop video which will hold the information of desired slices for each time point. The inputs for each dicom series image were the cine length which means the number of cine sequences to be stored in the cine loop, start and end slices considering the proper visualization of the region of interest which is the left ventricle in this case, and the step size which means the distance (in millimeters) between each slice in the cine loop sequence. After that, the dicom series 4D images will be transformed into a series of cine sequence videos which will hold the temporal information for all the slices. After extracting time points for each case, the next step is to extract slices for each time point for all the patient cases. A Matlab script was written to extract the video frames from each video setting the distance between the vertical slices. The start and end slices were chosen and then the picture frame was cropped so that each slice will hold the region of interest and not contain unnecessary pixels. After extracting the slices from each time point they are ready for image registration to get the deformation field. A set of slices as an example after the data preparation step is demonstrated in Figure 1.

3.2. Registration

The target of this step is to register the slices with respect to each time point to derive the deformation field. Each slice image at a particular time point (t_n) will be registered with respect to the initial time point (t_0) and the previous time point image (t_{n-1}). For performing image registration, *SimpleElastix* by Lowekamp et al. (2016) and *Cardiac motion estimation* library by Wiputra et al. (2020) were used. Here, the cardiac motion tracking uses the Fourier b-splines spatiotemporal motion model to fit the deformation fields. It requires the initial and final time points with the number of slices to be specified. After setting up the paths and initializing the bspline-solver, it performs the pairwise registration and stores the displacement fields for each pair with the scaling and transformation parameters. For each single

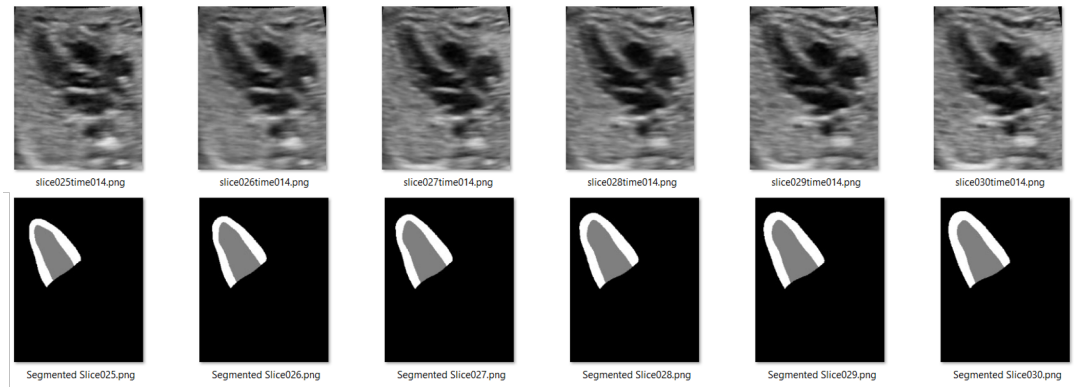


Figure 3: Sample intensity image slice and mask pairs.

image slice at (t_n), there will be two displacement fields which will be later combined using a weighted average to transform and derive the mask for that time point (t_n).

3.3. Segmentation

The next step after registration is manual segmentation or annotation. For this, two specific time points (preferably end-systolic/initial and end-diastolic/final) were chosen for each case. After that, all the slices for those time points were manually annotated. For performing the annotation, a quick and interactive segmentation technique called "*Lazy Snapping*" by Li et al. (2004) was chosen. It helps to choose the foreground and background by using a marker and based on that it generates the mask for each particular slice. The greater number of slices, the more robust data but also the number of slices might make the process a bit time-consuming as manual segmentation takes a considerable amount of time. After generating the segmentation mask, they were also checked and assessed by experts and their feedback was received. The generated segmentation masks would be irregular or not smooth enough as they might have staircase effects or holes. These will be corrected and smoothed in the later steps before generating the masks for other time points. A set of segmented masks as an example after the manual annotation using lazy snapping can be seen in Figure 2.

3.4. 3D Reconstruction

After generating the left ventricle masks for end-systolic/initial and end-diastolic/final time points, the next step is to combine these 2D slices to reconstruct the 3D mask for those points. For 3D reconstruction, "*VMTK (Vascular Modeling Toolkit)*" by Izzo et al. (2018) has been used which is a popular software for vascular image reconstruction and geometric analysis. The paths for all the slices were given as inputs and the result was the 3D reconstructed mask for the left ventricle at a given time point. As the results from the lazy snap step were not smooth and contained some artifacts, these 3D masks were corrected and smoothed with the help of an expert using "*Geomagic Wrap*" Software. This reverse engineering software helped to smoothen and regularize the mask by removing the artifacts. After the 3D mask was approved by the expert, later it was used to generate the other time point masks. To generate the 3D masks for other time points, the deformation fields obtained in the image registration step were used. Finally, for all the patient cases, 3D masks were generated for all the time points which were later used for training and testing the deep learning models. An example of the 3D mask can be seen in Figure 4.

3.5. Image Preprocessing

The intensity images acquired through the ultrasound scanner generated some artifacts like constant white boxes or arrows in the image which can be seen in Figure 5. For better performance during train, these constant regions should be removed or replaced by the neighboring pixel intensities as they might generate undesired results during training. As these artifacts were common and at the same position for all the images over slice and time for any cases, the same step for removing these artifacts from one image has been applied for all. To remove, the constant regions from the image, a simple linear interpolation method was used. In this method, an interpolation line was drawn between the left and the right pixel of the defected area, and then the defected area was interpolated using the intensity values from the interpolation line. The sample results can be seen in the same Figure 5.

3.6. Image-Mask Pair Generation

In the last step, the inner wall of the reconstructed masks was filled and reconstructed masks were binarised where (class 0 represents the background, 1 for the cavity of the left ventricle, and 2 for the myocardium. Later 3D masks were paired with the corresponding 3D intensity images for each time point to finalize the dataset for training. In the end, 14 4D echocardiography

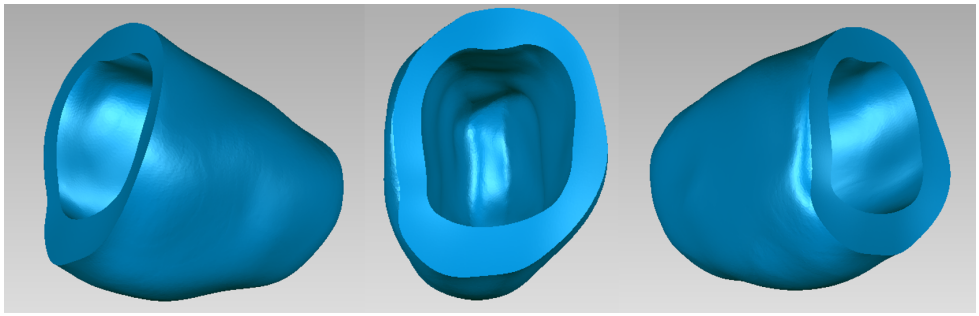


Figure 4: Example 3D Mask for Patient001 (Time014).

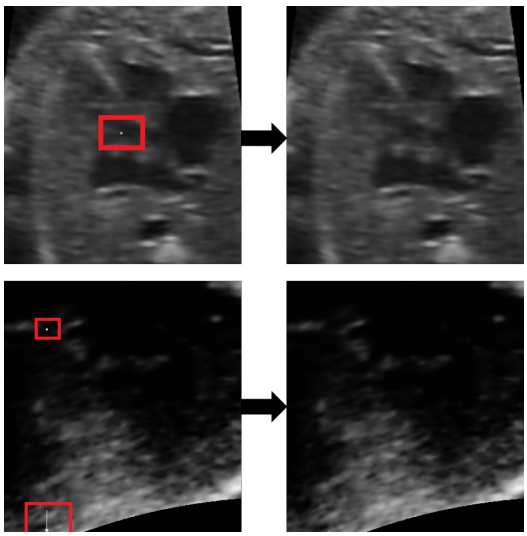


Figure 5: Intensity image artifact removal example.

images were transformed into a total of 518 3D images where each of the 3D images holds around 40 2D slices. As the nifty formatted files are hard to visualize in the report, a sample of slices for image and mask pairs are shown in Figure 3.

4. Experimental Methods

Let's assume f and m are two volume images that can be referred to as target/fixed image and moving image. The goal is to deform the moving image so that the anatomical location for all the voxels in fixed and moved images will be the same. Deep learning-based image registration (DLIR) neural networks will be used to model the displacement field which will transform all the voxels in the moving image so that they can be aligned with the fixed image. Let's say, the displacement field u will be modeled by CNN as the function $g_\theta(f, m) = \mathbf{d}$, where $\mathbf{d}$ is the displacement field and θ is the set of parameters learned by the CNN network. The main aim is to optimize the set of parameters so that the expected loss function can be minimized using Stochastic Gradient descent. Several approaches and experiments have been conducted to perform optimal image registration. The approaches will be discussed as follows.

4.1. Approach 1: Vanilla-DLIR

The underlying architecture of Vanilla DLIR is based on the traditional UNET architecture by Chen et al. (2021); Ronneberger et al. (2015) used for segmentation. The UNET consists of encoding and decoding layers with residual skip connections. This can be seen in Figure 6. The network used receives input fixed and moving images both of $256 * 256 * 32$ sizes which are concatenated to 2-channel 3D images. The 3D convolution is applied both in the encoding and decoding layers with a kernel size of 3, the stride is kept as 2 which is followed by Batch Normalization and ReLU layers. Max pooling is applied for downsampling in the encoding layers to reduce the spatial dimension of the image by half. The number of channels increases where the image size is reduced for the coarser representation of the input in the pyramid hierarchy. The bottleneck layer after the encoding layers captures the most abstract feature of the input image volume.

Then, the decoding layers perform the upsampling and convolutional operations to generate the displacement field. The convolutional layers consist of transposed 3D convolutions followed by batch normalization and ReLU layers. Skip connections from the encoding layers directly applied by concatenating. The conventional path cannot degrade the features' quality as a non-zero regularizing path will skip over them. On the other hand, the direct skipping of the non-zero regularizing path cannot hamper the performance as it has been added to the conventional path's learned features. Each layer of the decoding stage generates a finer spatial scaled image for generating the deformation field as an output of the final convolutional layer containing a 1*1 image filter and a softmax activation function.

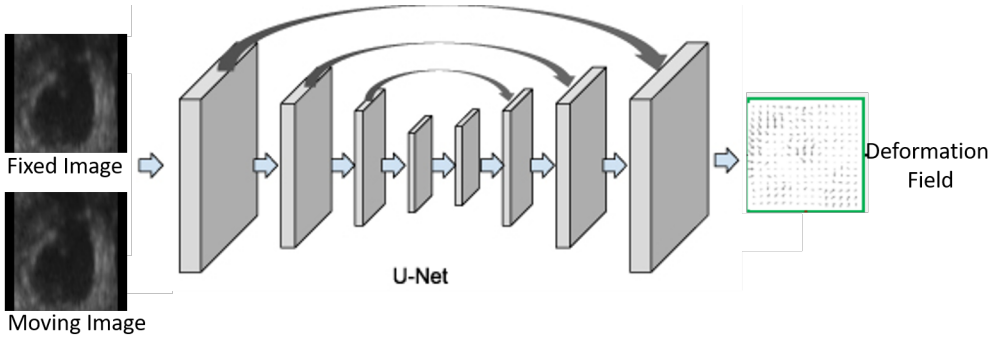


Figure 6: UNET for Image Registration with Skip connection.

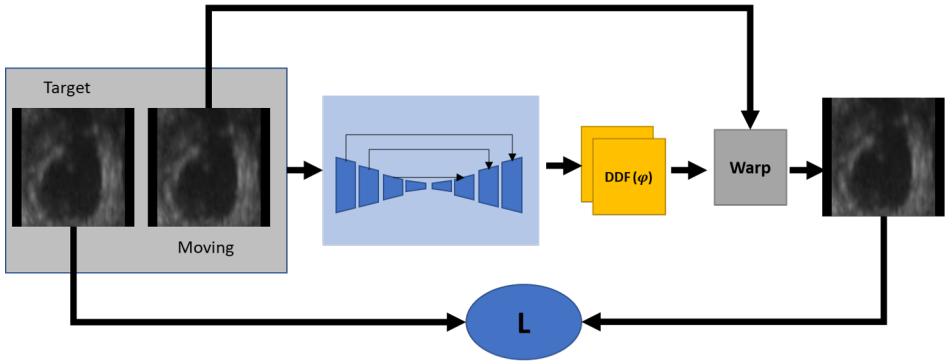


Figure 7: Vanilla-DLIR architecture.

4.1.1. *Vanilla-DLIR Loss Functions*

For vanilla-DLIR, unsupervised loss functions have been incorporated which consists of two components mainly. The first component is the similarity loss for having a better approximation of the fixed image in appearance for the moved image. Whereas, a regularization loss function called binding energy loss is used to penalize the non-regular spatial differences in order to have a smoother and more plausible displacement field. The equation for the total unsupervised loss is as follows where λ is a regularization parameter.

$$\mathcal{L}_{us}(f, m, d) = \mathcal{L}_{\text{sim}}(f, m \circ d) + \lambda \mathcal{L}_{\text{smooth}}(d) \quad (1)$$

There are a couple of similarity loss functions that can be used such as mean squared error(MSE), cross-correlation(CC), etc but for this work, Global Mutual Information(GMI) loss has been used. The statistical dependency or mutual information between two random variables, generally the fixed and the deformed moving image using the displacement field, is measured by the GMI loss. GMI loss seeks to maximize the similarity in appearance between the fixed image and the produced output. The model is compelled to acquire meaningful and instructive representations by maximizing mutual information. Firstly, the mutual information between the local patches is calculated and then the local patches MIs are aggregated to get the global mutual information. To calculate the mutual information for the patches in f and m, the following equation can be used,

$$I(f_x; m_y) = \sum_{x \in X} \sum_{y \in Y} P(f_x, m_y) \log_2 \left(\frac{P(f_x, m_y)}{P(f_x) P(m_y)} \right) \quad (2)$$

Higher mutual information yields a better alignment, so minimizing the negative GMI loss, the model tries to maximize the MI between the fixed and moved images.

GMI loss enforces the model to approximate the fixed image but the produced output may not be as smooth as desired. To have a smooth and more physically realistic deformed moving image, binding energy loss is also used in addition to GMI loss. Using a diffusion regularizer can leverage the spatial gradients of the deformation, u.

$$\mathcal{L}_{\text{smooth}}(d) = \sum_{\mathbf{d} \in D} \|\nabla \mathbf{u}(\mathbf{p})\|^2 \quad (3)$$

The differences between neighboring pixels in the 3D image are used to approximate the spatial gradient. The resulting architecture of Vanilla-DLIR with its loss function can be seen in Figure 7.

4.2. Approach 2: Anatomically Constrained DLIR

Anatomical masks of the Myocardium and left ventricle cavity are available from the data annotation part, the vanilla-DLIR can leverage from it. Balakrishnan et al. (2018) and Hu et al. (2017) in their respective research works showed that, the use of deformed segmentation masks during training enhances the performance of image registration in Vanilla-DLIR. In order to leverage the segmentation masks, first the registration field d, derived from the model network was used to deform the fixed image mask. After that, the segmented mask of the deformed image became available during training. As the segmented masks assign labels to the specific regions in the image, the same specific region in the fixed mask and deformed mask should also overlap. That was the key idea of getting the use of supervised loss in addition to the unsupervised loss for Vanilla-DLIR. Dice (1945) shows to quantify this volume overlap, Dice Score can be used. For example, the regions of either myocardium or left ventricle cavity, in this case, can be expressed in terms of the fixed and moved image can be expressed as $r_f^v and r_m^v \circ d$. The dice score can be computed to quantify the overlap of both regions as follows.

$$\text{Dice}\left(r_f^v, r_m^v \circ d\right) = 2 \cdot \frac{\left|r_f^v \cap (r_m^v \circ d)\right|}{\left|r_f^v\right| + |r_m^v \circ d|} \tag{4}$$

The dice score lies between 0 to 1, from no overlap to complete overlap. The dice score loss was defined $\mathcal{L}_{\text{dice}}$ over the whole segmented regions $v \in [1, V]$ as:

$$\mathcal{L}_{dice}\left(r_f, r_m \circ d\right) = -\frac{1}{K}\sum_{v=1}^{V} \text{Dice}\left(r_f^v, r_m^v \circ d\right) \tag{5}$$

4.2.1. Latent Space Consideration

In addition to dice score loss, the global anatomical constraint was also considered to compute the global loss. The local segmentation-aware loss computed by dice loss (fixed and moved labels) uses pixel-level predictions and may not ensure a satisfactory global match between the warped source and target anatomical masks shown by Oktay et al. (2017). Here, the segmentation masks for the fetal echo image volumes, represent the myocardium and left ventricle cavity. Segmentation masks represent pathological entities like brain tumors or skin lesions, which are very irregular in shape and topology. Whereas, Human organs like this scenario are highly regular, and are used to constrain registration. So, the plausibility of the shape is very important to get the correct registered images. For this reason, the latent space of the both target and the moved mask was considered to compute the global loss function. The global loss function considers the anatomical plausibility of the deformed source mask when comparing it to the target mask. Moreover, Oktay et al. (2017) also shows that, local dice loss acts at the pixel level, and back-propagated gradients are parametrized exclusively by pixel-wise individual probability components and provide little global context. To put global context in the loss computation, variational encoders were used to transform the target and moved masks to latent space, and compute global loss. The idea of a variational autoencoder can be understood in the next section and visualized in Figure 8.

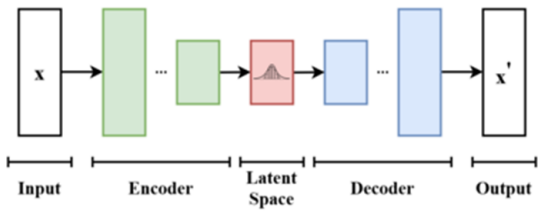


Figure 8: Learning global anatomical features.

4.2.2. Variational Autoencoder

To compute the global loss from the observations, the segmented masks needed to be transformed into latent space. A Variational autoencoder exactly does the same as shown by Oktay et al. (2017). Variational autoencoders(VAE) provide a probabilistic manner to describe the observations in latent space. In this work, the idea of VAE was adapted with a little bit of change in the architecture can be seen in Figure 9 to make it work for fetal echo masks for the myocardium and left ventricle. VAE contains two parts, encoders and decoders with a bottleneck layer. Encoders learn effective data encoding from datasets and pass it into bottleneck architectures. The autoencoder's decoder employs latent space in the bottleneck layer to generate dataset-like images. These results backpropagate from the neural network in the form of the loss function. For this work, the encoder part had 4 hierarchical stages each containing a block of convolutional neural network having kernel size=3 layer followed by batch normalization and ReLU layers. The downsampling was done by max pooling having kernel size=3, stride=2, and pooling=1. Residual connections were introduced at each stage to improve the flow of gradients during training. The inputs of the encoder were the single channel mask volumes of size $256 * 256 * 32$ which were halved at later stages. The bottleneck layer was a linear network transforming the output from the encoder to the latent space and passing it to the decoder. The decoder has the same 4 stages as the encoder where each stage has 3 blocks of convolutional neural network followed by batch normalization and ReLU.The upsampling was done with a scale factor

of 2 using trilinear interpolation. Finally, after the last stage, the top input-like images were reconstructed.

The loss functions for the variational autoencoders were a combination of 4 loss functions.

- Dice Score loss
- Euclidean L2 norm loss
- Structural Similarity loss
- Kullback-Leibler(KL) Loss

The dice score loss is computed between the input and reconstructed image using the equation 5.

The Euclidean L2 norm loss computes the Euclidean distance between the input images and the reconstructed ones. Let's say, if i and r are the input and reconstructed masks respectively, the L2 loss was computed by the following equation:

$$\mathcal{L}_{L2}(i,r) = \frac{1}{N}\sum_{n=1}^{N}(i_n - r_n)^2 \tag{6}$$

L2 norm penalizes the larger distances between the voxels in input and reconstructed masks more than the smaller distances.

To assess the quality of the image reconstruction by guiding the image generation, the structural similarity measure index was also computed as shown by Wang et al. (2004). Structural similarity loss can be computed to penalize the dissimilarity between the input and the reconstructed masks. The following equation was used to compute the SSIM loss:

$$\mathcal{L}_{SSIM}(i,r) = 1 - \frac{(2\mu_i\mu_r + C_1)(2\sigma_{ir} + C_2)}{\left(\mu_i^2 + \mu_r^2 + C_1\right)\left(\sigma_i^2 + \sigma_r^2 + C_2\right)} \tag{7}$$

where μ_i and μ_r are the average pixel intensities of i and r, σ_i and σ_r are the standard deviations of pixel intensities. Finally, σ_{ir} is the covariance between the pixel intensities of the two images. C1 and C2 are small constants added to stabilize the division when the denominator approaches zero.

The regularization loss named Kullback-Leibler (KL) divergence in Kingma and Welling (2014) forces the distributions returned by the encoder to be close to a standard normal distribution. KL loss will be a good representative to assess the discrepancy between the latent and desired distribution, and thus in generative models like VAEs, the KL divergence can be often used as a regularization term. The goal is to penalize the discrepancy between the learned latent distribution and a prior standard normal distribution. Let's say, for the standard normal distribution prior is $P(z)$, and the learned approximate posterior $Q(z|x)$, KL loss will be:

$$\mathcal{L}_{KL}(P(z), Q(z|x)) = \frac{1}{2}\sum\left(\mu^2 + \sigma^2 - \log(\sigma^2) - 1\right) \tag{8}$$

where μ and σ are the mean and standard deviation of the approximate posterior distribution $Q(z|x)$ for each latent variable z and will be summed for all latent variables. Finally, the variational autoencoder is trained to optimize the total loss function which can be described as:

$$\begin{aligned}\mathcal{L}_{va}(i,r,P(z),Q(z|x)) = \mathcal{L}_{dice}(i,r) + \mathcal{L}_{L2}(i,r) \\ + \mathcal{L}_{SSIM}(i,r) + \mathcal{L}_{KL}(P(z),Q(z|x))\end{aligned} \tag{9}$$

For training and validating the variational autoencoder, out of 518 3D annotated volume masks discussed in the dataset description section, 452 volume masks were used for training and the rest for validation. To improve the generalization of VAE, some data augmentation techniques like flipping and center-cropping were also used. An example of the results after the training of VAE can be seen in Figure 11.

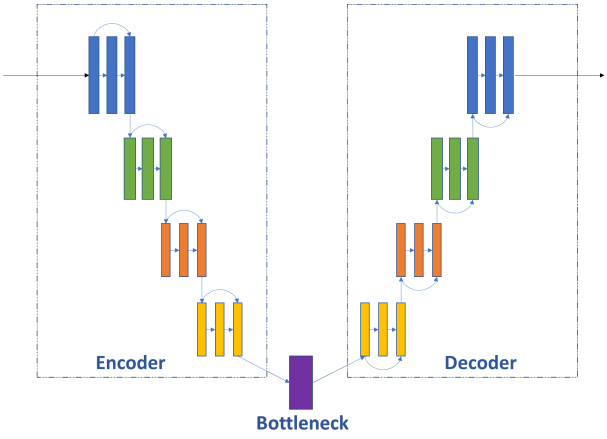


Figure 9: Variational Autoencoder Architecture.

4.2.3. AC-DLIR Loss Functions

The unsupervised loss introduced for vanilla-DLIR and the dice score loss from equation 5 are combined. In addition to that, image global loss is computed too. For computing global loss, the latent space consideration from VAE is used. Both the input and predicted mask are passed by the variational autoencoder model to generate the reconstructed masks. The global loss is computed between these two reconstructed masks both for the myocardium and left ventricle and added together. The global loss is the computation of the L2 norm which is discussed in equation 6. The total loss with anatomical constraint consideration for AC-DLIR is:

$$\begin{aligned}\mathcal{L}_a\left(f,m,r_f,r_m,d\right) = \mathcal{L}_{us}(f,m,d) \\ +\beta\mathcal{L}_{dice}\left(r_f, r_m \circ d\right) + \gamma\mathcal{L}_{L2}(r_f,r_m)\end{aligned} \tag{10}$$

where, both β and γ are regularization parameters. Fi-

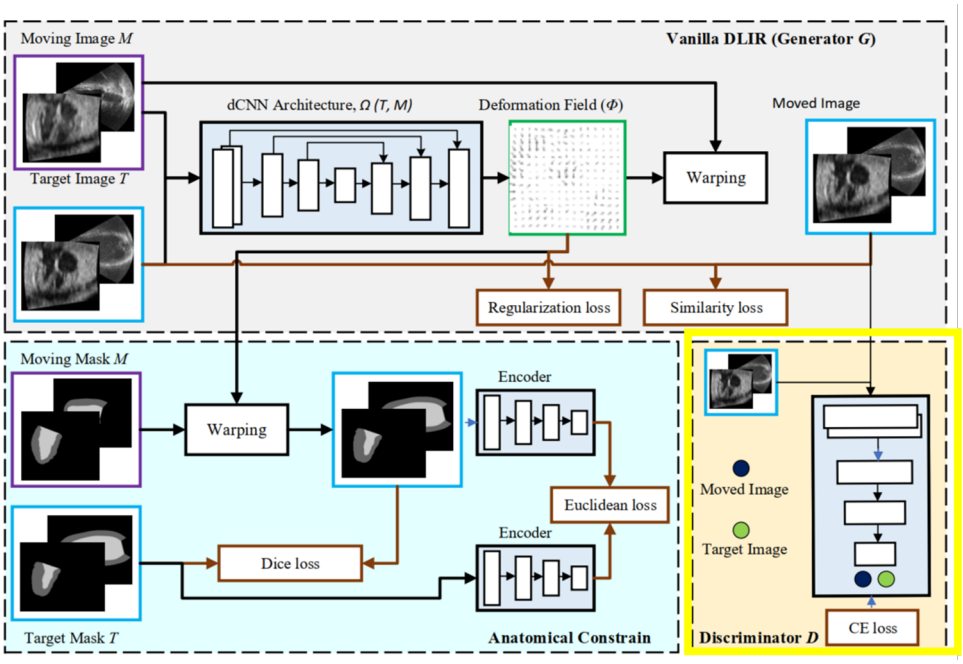


Figure 10: Proposed Adversarial Anatomically Constrained (AdvAC) DLIR architecture.

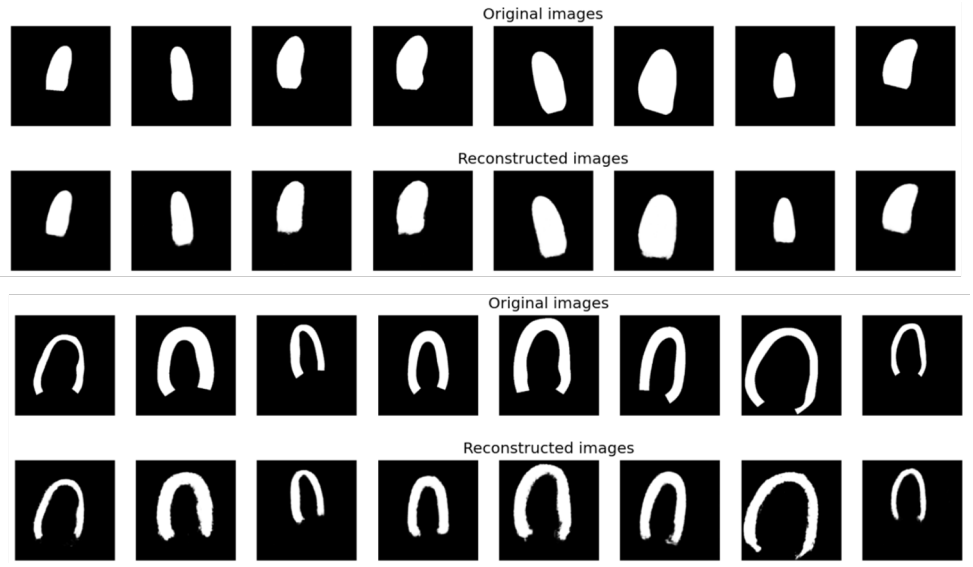


Figure 11: Original and reconstructed masks by VAE (top: left ventricle, bottom: myocardium).

nally, the architecture used for AC-DLIR can be visualized in Figure 10 excluding the highlighted part.

4.3. Approach 3: Adversarial AC-DLIR

The next addition to the network proposed is the inclusion of adversarial learning. As shown by Mahapatra et al. (2018), the use of the GAN network as a zero-sum game theory could be beneficial for learning deformable fields in image registration. In the proposed network, the part of AC-DLIR for generating the deformable images with the produced deformation field was treated as a generator for the adversarial network. In addition to that, a discriminator was also created which was able to classify the fixed and moved images. The architecture of the discriminator consists of 5 layers each containing convolutional blocks with 2 residual units outputting 8,16,32,64 and single channels respectively. The input was the single channel input image volume. Kernel size was kept at 3 with strides 2,2,2,2 and 1 at the respective layers and with LeakyReLU activation. The dropout layer was also used with a probability of 0.10. For the loss function of both the generator and discrim-

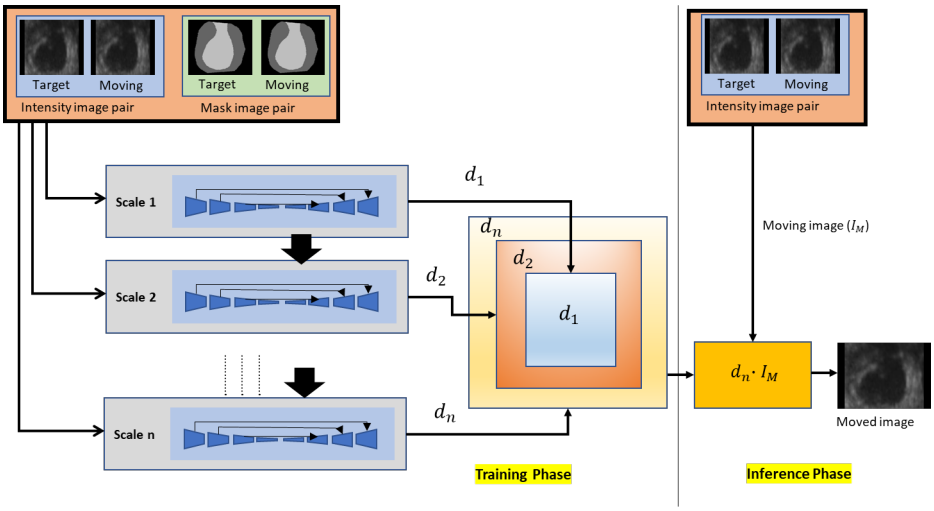


Figure 12: Proposed MACMR architecture.

Models	MSE	Dice Score			Mean Dice ± std
		BG	Myo	LV	
Without Registration	0.00972	0.96678	0.69391	0.76046	0.80235±0.05491
Vanilla-DLIR	0.0042	0.97352	0.74977	0.87523	0.88487±0.03261
AC-DLIR	0.00598	0.97972	0.81437	0.91935	0.90303±0.03447
Adv-DLIR	0.00533	0.97429	0.79278	0.86842	0.85733±0.04129
AdvAC-DLIR	**0.00589**	**0.98742**	**0.82751**	**0.93573**	**0.91689±0.02596**
MACMR	**0.00489**	**0.98779**	**0.84871**	**0.95423**	**0.94245±0.02474**

Table 1: Comparison of proposed registration models on CAMUS 2D Dataset.

Models	MSE	Dice Score			Mean Dice ± std
		BG	LV	Myo	
Without Registration	0.00377	0.99093	0.78917	0.72605	0.83539±0.12798
Vanilla-DLIR	0.00296	0.98699	0.70087	0.58543	0.75776±0.04036
AC-DLIR	0.00251	0.98959	0.73347	0.64435	0.80013±0.05401
Adv-DLIR	**0.00339**	**0.99031**	**0.73836**	**0.67389**	**0.80989±0.05142**
AdvAC-DLIR	**0.00258**	**0.99089**	**0.79884**	**0.73482**	**0.84668±0.04586**

Table 2: Comparison of proposed registration models on Fetal 3D Dataset.

inator, the binary cross-entropy loss was used. While training, the generator, and discriminator will fight o gain over each other as the task of the generator would be creating as much as plausible images as the fixed image whereas the discriminator would try to discriminate them. The loss from the generator was added to the $\mathcal{L}_a(f, m, r_f, r_m, d)$ from equation 10 as the deformable field generated by training would be capable of better generalization if the loss of the generator was being optimized.

$$\mathcal{L}_{adac}\left(f, m, r_f, r_m, d, s_m\right) = \mathcal{L}_{us}(f, m, d) + \beta\mathcal{L}_{dice}\left(r_f, r_m \circ d\right) + \gamma\mathcal{L}_{L2}(r_f, r_m) + \phi\mathcal{L}_g(m, s_m) \quad (11)$$

In this loss function, equation, ϕ is a regularization parameter set as 0.0001, m and s_m are the moved image and assigned real labels to the moved image. The final architecture after adding the adversarial network to the AC-DLIR can be seen in figure 10.

4.4. Approach 4: Multi-Scale Registration (MACMR)

The final proposal to improve the performance of image registration is Multi-scale (multi-resolution) training, where trained parameters on the lower scale will be used to initialize the higher-scale training. As the features learned at the lower scales can guide the training for the higher scale, the network at a higher scale will have a better initialization. Better initialization of the network should result help the network converge faster to achieve better performance. Moreover, it can be shown that Multi-resolution training helps the network to learn both local and global information. It can improve the performance of the model with various

scales and enhance its overall performance. The proposed MACMR architecture is demonstrated in Figure 12.

5. Results

We have implemented all the methods discussed above in Pytorch. The experiments were performed on NVIDIA GeForce RTX 3090 Ti. The inputs were kept as 256*256*32 resolution for the fetal dataset. We have performed an analysis of the performance of the model for both 2D Camus and 3D Fetal datasets. During the experiments, the Adam optimization technique was used and the learning rate was kept at 0.001 with the use of a learning rate scheduler. We have trained the model with 100 epochs. For the training of variational autoencoders, the same hyperparameters were used with 200 epochs. For evaluation and comparison of the results, we have used mean squared error from equation 6 and Dice Score Coefficient from equation 5 were used.

The detailed comparison between the proposed models can be seen in Table 1 and 2. We have also visualized the results for 2D slices which can be seen in the appendix from figure 13, 14, 15, 16. In the figure, the masks were colored according to the overlapping of the pixels where green means true positive, and yellow and red define pixels which are false positive and false negative. The fifth column indicates the overlap of fixed and moving images whereas the sixth column indicates the overlap of the fixed and moved image slices.

6. Discussion

The results are presented for two datasets: 2D CAMUS and 3D Fetal in 1 and 2 respectively. The computation of evaluation metrics between fixed and moving images is referred to as without registration. After registration, evaluation metrics are again computed between fixed and deformed images. The first experiment was done using the baseline model Vanilla-DLIR. We can see that the mean-squared error decreases after registration which indicates that in the case of vanilla-DLIR, the unsupervised registration without considering the anatomy, the similarity between two intensity images increases, but the similarity between fixed and moved masks does not improve satisfactorily or fail in some cases. For that reason, the DSC of the left ventricle and myocardium does not improve much. Figure 13 also shows that the overlapping of the fixed and deformed images is highly irregular containing false positive and negative cases.

Next, we tried to add latent space training to extract the global features using variational autoencoders. In this experiment, we can see the MSE metric decreases as well as the DSC improves than Vanilla-DLIR. Figure 14 also indicates a better overlapping. As VAEs add global context to the learning, model, the results also prove that adding global latent space learning can be beneficial to perform better registration.

The overlapping in the images shows that the boundaries of the regions segmented are irregular or not very smooth. In the third experiment, we tried to add adversarial learning to provide better regularization of the model. From the results both from the table and the images, it can be seen that adversarial learning provides a better regularization and thus also improves the result of vanilla-DLIR.

So, we decided to keep them both in the model and apply them to perform the registration. The results of the proposed AdvAC model outperforms all the previous experiments and thus proved to be the best model working in both the 2D and 3D dataset. Still, there is room for performance improvement. Still, there is room for performance improvement. Hence, we proposed Multi-class Anatomically Constrained and Multi-scale Registration (MACMR) framework which is the best-performing model for the 2D Camus dataset. Although the results on 2D dataset is higher but both 2D and 3D data have the same upward improvement with the proposed models. The fact is that the volume images are low in number for training and also take longer time than 2D for training for each epoch, the result is lower but still satisfactory as this will be the first time temporal registration was done on 3D fetal echocardiography images. In our future plan, we want to add even more 3D data volume to have a better training of the model and also want to apply the multi-resolution framework in case of 3D.

7. Conclusions

The clinical use of echo is still stuck with 2D, likely because doctors can not visualize 3D, but for machine learning it makes more sense to go 3D, for real-time detection with improved accuracy and precision. Existing DLIR or DL echo image processing are all 2D, and so the need for 3D temporal registration for echo images is clearly visible. Also there is less research work done for fetal hearts although the fetal heart can experience congenital heart malformation and functional abnormalities. This thesis focuses on the development of methods for automatic 3D temporal registration for 3D fetal echocardiographic images. The aim was to improve the detection of congenital heart malformations and functional abnormalities in the developing fetus.

One of the two most important aspects of this thesis was to propose a new dataset for fetal echocardiography. 4D volume echocardiography images were collected and annotated with the use of a cardiac motion estimation algorithm. We have conducted several experiments starting with UNET-based DLIR to adding global latent space training with variational autoencoders and adversarial learning to have a better regularization loss.

We have compared the results for both 2D and 3D datasets. The results have shown significant improvements in temporal registration accuracy using evaluation metrics such as Mean Squared Error, and Dice Metric. As the data annotation takes a considerable amount of time, we started the work with a few number of volume images which hindered the overall performance of the 3D dataset. So, we are planning to add more annotated data as well as to evaluate the 3D model in multi-resolution framework.

References

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J.V., Dalca, A.V., 2018. Voxelmorph: A learning framework for deformable medical image registration. IEEE Transactions on Medical Imaging 38, 1788–1800.

Chen, X., Diaz-Pinto, A., Ravikumar, N., Frangi, A.F., 2021. Deep learning in medical image registration. Progress in Biomedical Engineering 3, 012003. URL: https://dx.doi.org/10.1088/2516-1091/abd37c, doi:10.1088/2516-1091/abd37c.

Cheng, X., Zhang, L., Zheng, Y., 2018. Deep similarity learning for multimodal medical images. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 6, 248 – 252.

Ciancarella, P., Ciliberti, P., Santangelo, T.P., Secchi, F., Stagnaro, N., Secinaro, A., 2020. Noninvasive imaging of congenital cardiovascular defects. La radiologia medica 125, 1167 – 1185.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Dong, J., Liu, S., Wang, T., 2019. Arvbnet: Real-time detection of anatomical structures in fetal ultrasound cardiac four-chamber planes, in: MLMECH/CVII-STENT@MICCAI.

Fechter, T., Baltas, D., 2019. One-shot learning for deformable medical image registration and periodic motion tracking. IEEE Transactions on Medical Imaging 39, 2506–2517.

Ferrante, E., Dokania, P.K., Silva, R.M., Paragios, N., 2018. Weakly supervised learning of metric aggregations for deformable image registration. IEEE Journal of Biomedical and Health Informatics 23, 1374–1384.

Fu, Y., Lei, Y., Wang, T., Higgins, K.A., Bradley, J.D., Curran, W.J., Liu, T., Yang, X., 2020. Lungregnet: an unsupervised deformable image registration method for 4d-ct lung. Medical physics .

Gong, L., Wang, H., Peng, C., Dai, Y., Ding, M., Sun, Y., Yang, X., Zheng, J., 2017. Non-rigid mr-trus image registration for image-guided prostate biopsy using correlation ratio-based mutual information. BioMedical Engineering OnLine 16.

Green, L., Chan, W.X., Ren, M., Mattar, C.N.Z., Lee, L.C., Yap, C.H., 2023. The dependency of fetal left ventricular biomechanics function on myocardium helix angle configuration. Biomechanics and modeling in mechanobiology 22, 629—643. URL: https://europepmc.org/articles/PMC10097781, doi:10.1007/s10237-022-01669-z.

Haskins, G., Kruecker, J., Kruger, U., Xu, S., Pinto, P.A., Wood, B.J., Yan, P., 2018. Learning deep similarity metric for 3d mr-trus registration. ArXiv abs/1806.04548.

Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T.N., Gleeson, F.V., Brady, M., Schnabel, J.A., 2012. Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. Medical image analysis 16 7, 1423–35.

Hu, Y., Modat, M., Gibson, E., Ghavami, N., Bonmati, E., Moore, C.M., Emberton, M., Noble, J.A., Barratt, D.C., Vercauteren, T.K.M., 2017. Label-driven weakly-supervised learning for multimodal deformarle image registration. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) , 1070–1074.

Izzo, R., Steinman, D.A., Manini, S., Faggiano, E., Antiga, L., 2018. The vascular modeling toolkit: A python library for the analysis of tubular structures in medical images. J. Open Source Softw. 3, 745.

Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K., 2015. Spatial transformer networks, in: NIPS.

Jafari, M., Girgis, H., Abdi, A.H., Liao, Z., Pesteie, M., Rohling, R.N., Gin, K., Tsang, T., Abolmaesumi, P., 2019. Semi-supervised learning for cardiac left ventricle segmentation using conditional deep generative models as prior. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) , 649–652.

Jafari, M.H., Girgis, H., Liao, Z., Behnami, D., Abdi, A., Vaseli, H., Luong, C., Rohling, R., Gin, K., Tsang, T., Abolmaesumi, P., 2018. A unified framework integrating recurrent fully-convolutional networks and optical flow for segmentation of the left ventricle in echocardiography data, in: Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J.M.R., Bradley, A., Papa, J.P., Belagiannis, V., Nascimento, J.C., Lu, Z., Conjeti, S., Moradi, M., Greenspan, H., Madabhushi, A. (Eds.), Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer International Publishing, Cham. pp. 29–37.

Kingma, D.P., Welling, M., 2014. Auto-encoding variational bayes, in: International Conference on Learning Representations (ICLR).

Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., Lartizien, C., D'hooge, J., Lovstakken, L., Bernard, O., 2019. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Transactions on Medical Imaging 38, 2198–2210. doi:10.1109/TMI.2019.2900516.

Li, Y., Sun, J., Tang, C.K., Shum, H., 2004. Lazy snapping. ACM SIGGRAPH 2004 Papers .

Lowekamp, B., gabehart, Blezek, D., Marstal, K., Ibanez, L., Chen, D., McCormick, M., Mueller, D., Johnson, H., Cole, D., Yaniv, Z., Posthuma, J., Beare, R., Gelas, A., aghayoor, ltong1130ztr, fsantini, adizhol, Subburam, K., Fillion-Robin, J.C., Anthony, Doria, D., King, B., 2016. kaspermarstal/simpleelastix: v0.10.0. URL: https://doi.org/10.5281/zenodo.168078, doi:10.5281/zenodo.168078.

Mahapatra, D., Antony, B.J., Sedai, S., Garnavi, R., 2018. Deformable medical image registration using generative adversarial networks. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) , 1449–1453.

Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M.P., Bai, W., Caballero, J., Cook, S.A., de Marvao, A., Dawes, T.J.W., O'Regan, D.P., Kainz, B., Glocker, B., Rueckert, D., 2017. Anatomically constrained neural networks (acnns): Application to cardiac image enhancement and segmentation. IEEE Transactions on Medical Imaging 37, 384–395.

Ong, C.W., Ren, M., Wiputra, H., Mojumder, J., Chan, W.X., Tulzer, A., Tulzer, G., Buist, M.L., Mattar, C.N.Z., Lee, L.C., Yap, C.H., 2020. Biomechanics of human fetal hearts with critical aortic stenosis. Annals of Biomedical Engineering 49, 1364 – 1379.

Painchaud, N., Skandarani, Y., Judge, T., Bernard, O., Lalande, A., Jodoin, P.M., 2019. Cardiac segmentation with strong anatomical guarantees. IEEE Transactions on Medical Imaging 39, 3703–3713.

Rivaz, H., Karimaghaloo, Z., Fonov, V.S., Collins, D.L., 2014. Nonrigid registration of ultrasound and mri using contextual conditioned mutual information. IEEE Transactions on Medical Imaging 33, 708–725.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. ArXiv abs/1505.04597.

Sachdeva, S., Gupta, S., 2020. Imaging modalities in congenital heart disease. The Indian Journal of Pediatrics 87, 385–397.

Simonovsky, M., Gutiérrez-Becker, B., Mateus, D., Navab, N., Komodakis, N., 2016. A deep metric for multimodal registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image

quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing 13, 600–612.

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

Wu, G., Kim, M., Wang, Q., Munsell, B.C., Shen, D., 2016. Scalable high-performance image registration framework by unsupervised deep feature representations learning. IEEE Transactions on Biomedical Engineering 63, 1505–1516.

Xie, H., Lei, Y., Fu, Y., Wang, T., Roper, J.R., Bradley, J.D., Patel, P.R., Liu, T., Yang, X., 2022. Inter-fraction deformable image registration using unsupervised deep learning for cbct-guided abdominal radiotherapy. Physics in Medicine and Biology 68.

Yan, P., Xu, S., Rastinehad, A.R., Wood, B.J., 2018. Adversarial image registration with application for mr and trus image fusion, in: MLMI@MICCAI.

Yang, M., Xiao, X., Liu, Z., Sun, L., Guo, W., zhen Cui, L., Sun, D., Zhang, P., Yang, G., 2020. Deep retinanet for dynamic left ventricle detection in multiview echocardiography classification. Sci. Program. 2020, 7025403:1–7025403:6.

Yoon, Y.E., Kim, S., Chang, H.J., 2021. Artificial intelligence and echocardiography. Journal of Cardiovascular Imaging 29, 193 – 204.

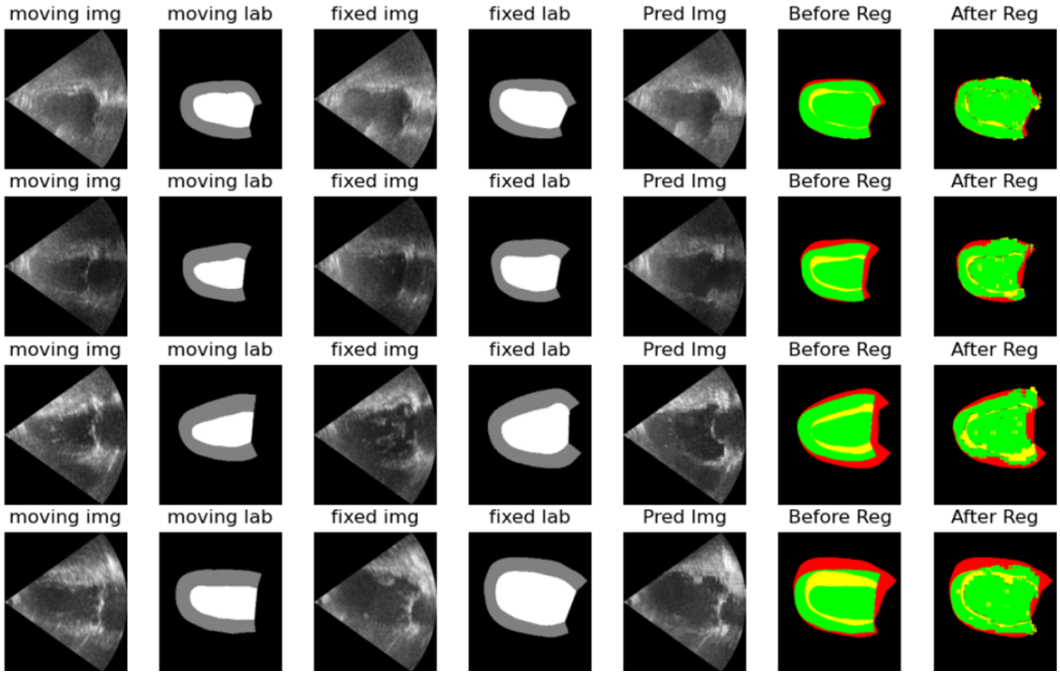


Figure 13: Segmentation Results for Vanilla-DLIR.

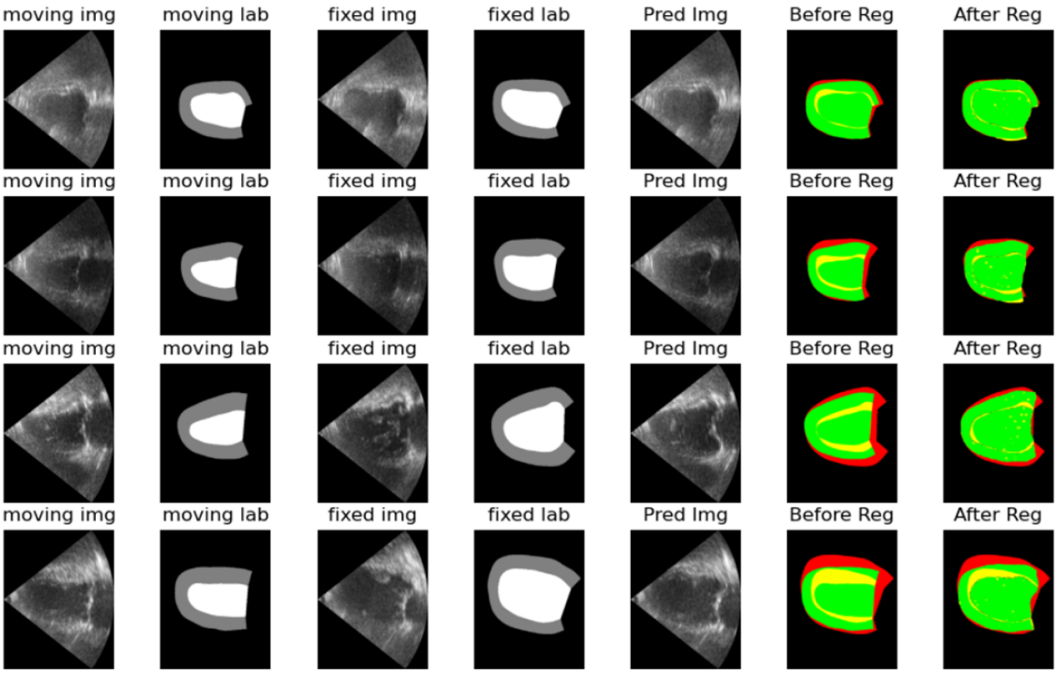


Figure 14: Segmentation Results for AC-DLIR.

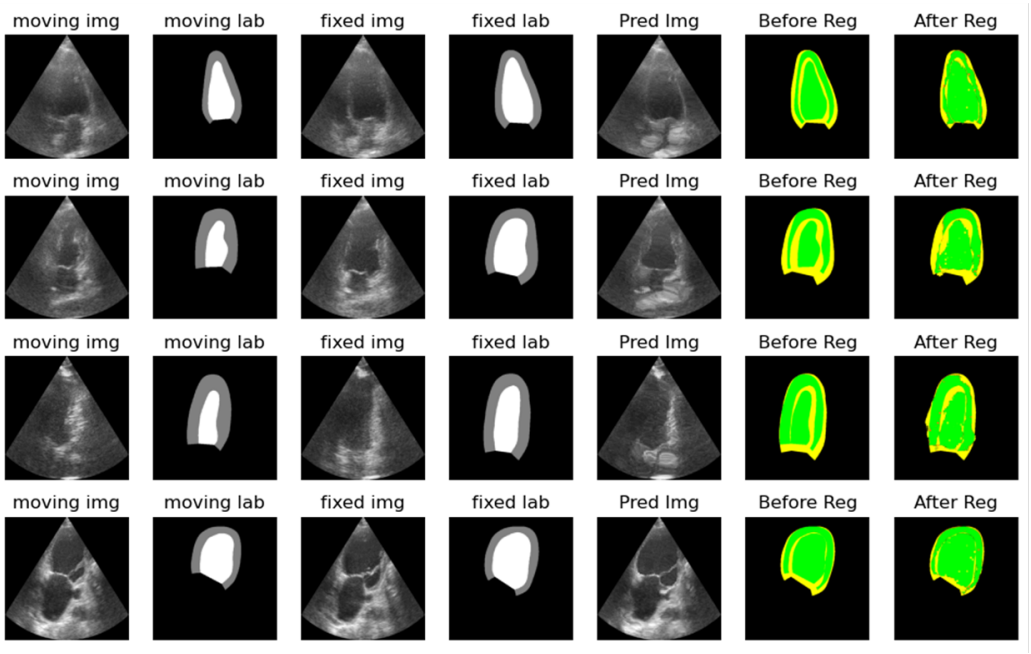


Figure 15: Segmentation Results for Adv-DLIR.

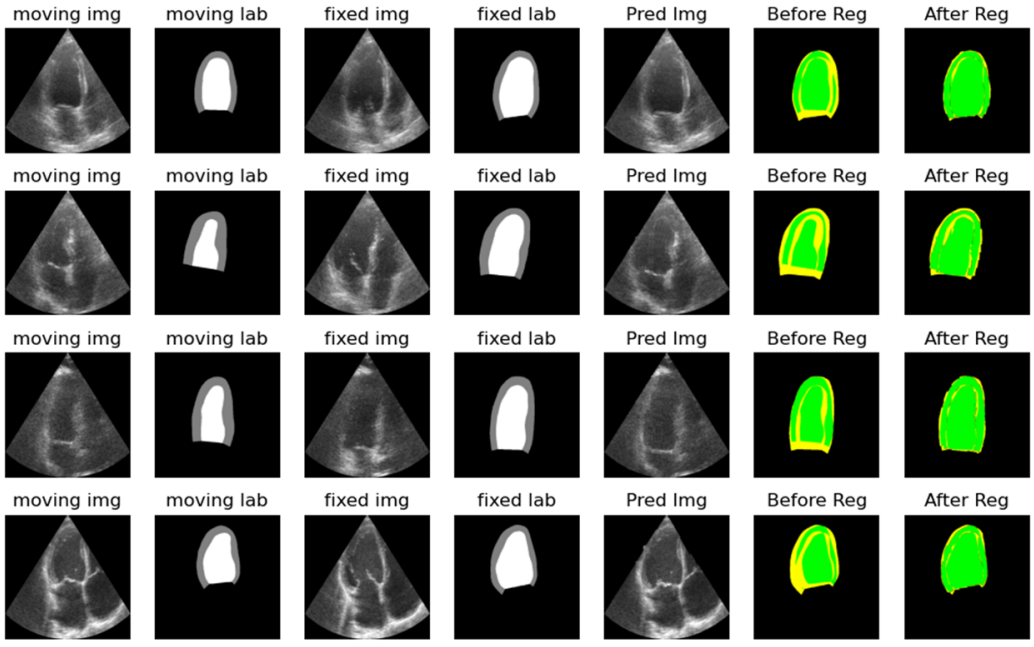


Figure 16: Segmentation Results for AdvAC-DLIR.