Medical Image Analysis 83 (2023) 102678

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

DragNet: Learning-based deformable registration for realistic cardiac MR sequence generation from a single frame

Arezoo Zakeri [a,1,*], Alireza Hokmabadi [a,1], Ning Bi [a], Isuru Wijesinghe [b], Michael G. Nix [c], Steffen E. Petersen [d,e,f,g], Alejandro F. Frangi [a], Zeike A. Taylor [b], Ali Gooya [g,h,**]

[a] Centre for Computational Imaging and Simulation Technologies in Biomedicine, School of Computing, University of Leeds, UK
[b] Centre for Computational Imaging and Simulation Technologies in Biomedicine, School of Mechanical Engineering, University of Leeds, UK
[c] Leeds Cancer Centre, Leeds Teaching Hospitals NHS Trust, UK
[d] William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University of London, UK
[e] Barts Heart Centre, St Bartholomew's Hospital, Barts Health NHS Trust, London, UK
[f] Health Data Research UK, London, UK
[g] Alan Turing Institute, London, UK
[h] School of Computing Science, University of Glasgow, Glasgow, UK

ARTICLE INFO

Keywords:
Deformable temporal image registration
Sequential image data generation
Deep learning
Variational recurrent neural networks
Uncertainty estimation
UK Biobank

ABSTRACT

Deformable image registration (DIR) can be used to track cardiac motion. Conventional DIR algorithms aim to establish a dense and non-linear correspondence between independent pairs of images. They are, nevertheless, computationally intensive and do not consider temporal dependencies to regulate the estimated motion in a cardiac cycle. In this paper, leveraging deep learning methods, we formulate a novel hierarchical probabilistic model, termed DragNet, for fast and reliable spatio-temporal registration in cine cardiac magnetic resonance (CMR) images and for generating synthetic heart motion sequences. DragNet is a variational inference framework, which takes an image from the sequence in combination with the hidden states of a recurrent neural network (RNN) as inputs to an inference network per time step. As part of this framework, we condition the prior probability of the latent variables on the hidden states of the RNN utilised to capture temporal dependencies. We further condition the posterior of the motion field on a latent variable from hierarchy and features from the moving image. Subsequently, the RNN updates the hidden state variables based on the feature maps of the fixed image and the latent variables. Different from traditional methods, DragNet performs registration on unseen sequences in a forward pass, which significantly expedites the registration process. Besides, DragNet enables generating a large number of realistic synthetic image sequences given only one frame, where the corresponding deformations are also retrieved. The probabilistic framework allows for computing spatio-temporal uncertainties in the estimated motion fields. Our results show that DragNet performance is comparable with state-of-the-art methods in terms of registration accuracy, with the advantage of offering analytical pixel-wise motion uncertainty estimation across a cardiac cycle and being a motion generator. We will make our code publicly available.

1. Introduction

Spatio-temporal motion tracking has been widely used in medical applications such as motion management in radiation therapy, tumour localisation, treatment planning, assessing organ motion in different image modalities (Giger et al., 2018; Teng et al., 2021; Mezheritsky et al., 2022), as well as analysing the heart motion along a cardiac cycle via motion indices (De Craene et al., 2012; Rohé et al., 2018; Krebs et al., 2019b, 2021). One way of motion tracking is to perform image registration which finds an optimal spatial transformation that best aligns two or more images based on some image similarity metrics. Traditional deformable registration algorithms iteratively solve an optimisation problem for each image pair. These techniques are computationally intensive and not applicable to real-time motion analysis. For real-time settings, the registration must be fast and accurate.

* Correspondence to: Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), School of Computing, University of Leeds, Leeds, UK.
** Corresponding author at: School of Computing Science, University of Glasgow, Glasgow, UK.
E-mail addresses: a.zakeri@leeds.ac.uk (A. Zakeri), Ali.Gooya@glasgow.ac.uk (A. Gooya).
[1] The first two authors contributed equally.

https://doi.org/10.1016/j.media.2022.102678
Received 8 March 2022; Received in revised form 24 August 2022; Accepted 27 October 2022
Available online 2 November 2022

Computation of the motion fields in a sequence is, therefore, more challenging using traditional techniques.

Deep learning (DL) based techniques have provided the means to significantly speed up the registration process for unseen images by proposing a trained network (Chen et al., 2021). These techniques can be categorised into supervised or unsupervised forms. The supervised methods rely on ground-truth displacement vector fields (DVFs), which are often provided by random transformation generation (Salehi et al., 2018; Eppenhof and Pluim, 2018; Eppenhof et al., 2018), conventional registration methods (Sentker et al., 2018; Fan et al., 2019), or model-based DVF generation techniques (Uzunova et al., 2017; Sokooti et al., 2019). Random transformations are generally different from the true physiological motion, which results in bias in the training and performance degradation. Sokooti et al. (2019) have shown that training a supervised model for 3D-CT lung image registration using a realistic model-based DVF generation performs better compared to the random transformations in terms of registration error. In most medical applications, the lack of training datasets with known ground-truth DVFs limits the utility of the supervised registration algorithms. Besides, lacking availability of training images of a certain kind is a big challenge. Uzunova et al. (2017) proposed a model-based data augmentation scheme to allow for deep learning on small training populations. They adapted the supervised FlowNet (Dosovitskiy et al., 2015) architecture for convolutional neural networks (CNN)-based optical flow estimation in cardiac images. This approach is limited to generating a diverse set of training image pairs with known correspondences but is not suitable for generating realistic sequential image data.

Unsupervised motion estimation techniques have effectively alleviated the data associated challenges of the supervised models. In these techniques, generally, some 2D or 3D convolutional layers are followed by a spatial transformer layer to form the structure of the unsupervised DIR networks (Balakrishnan et al., 2019; Dalca et al., 2018; Krebs et al., 2018, 2021; De Vos et al., 2017, 2019; Kuang and Schmah, 2019). Jaderberg et al. (2015) proposed the spatial transformer network (STN), which has been utilised frequently as a component in the unsupervised registration models (Balakrishnan et al., 2019; Dalca et al., 2018; Krebs et al., 2018, 2021; De Vos et al., 2017). STN deforms the moving image in a fully differentiable manner and enables image similarity optimisation during training. Balakrishnan et al. (2019) proposed an unsupervised CNN-based model in the UNet structure, termed VoxelMorph, for brain MRI registration. VoxelMorph achieves comparable performance to the non-learning-based methods such as ANTs SyN (Avants et al., 2008, 2011) and NiftyReg in terms of the Dice score of multiple anatomical structures while operating orders of magnitude faster (Balakrishnan et al., 2019). De Vos et al. (2017) proposed the 2D DIRNet model consisting of a CNN regressor, a spatial transformer, and a resampler and tested that on MNIST and short-axis (SAX) cardiac MR (CMR) image slices. Later, the 3D deep learning image registration (DLIR) framework consisting of a stack of CNNs was proposed for multi-stage unsupervised affine and deformable image registration (De Vos et al., 2019). The model was evaluated on image pairs of cardiac MRIs and chest CT with comparable results to the conventional SimpleElastix method (Marstal et al., 2016), but achieved a faster running time (De Vos et al., 2019). FAIM, a CNN model for 3D Brain MR image registration, includes a penalty loss on negative Jacobian determinants to decrease regions of non-invertibility (Kuang and Schmah, 2019). With the same objectives, Zhang (2018) proposed Inverse-Consistent deep Network (ICNet) on T1-weighted brain MRI, which controls the diffeomorphic property of the transformation by inverse-consistent and anti-folding constraints.

These learning-based DIR models, although valuable in many aspects, cannot generate synthetic motion beyond the registration task. This is interesting for recovering missing frames in a sequence, data augmentation, and even validating supervised DIR algorithms. Recently, probabilistic frameworks were suggested for this purpose (Dalca et al., 2018; Krebs et al., 2018, 2019a,b, 2021). Dalca et al. (2018) proposed a probabilistic model based on a 3D UNet-style architecture and applied that on brain MRIs enforcing diffeomorphic registration by introducing scaling and squaring differentiable layers. Krebs et al. (2018, 2019a) proposed a low-dimensional probabilistic parameterisation of deformations using a conditional variational autoencoder (CVAE) network. They utilised a Gaussian smoothness kernel followed by a differentiable exponentiation layer to obtain diffeomorphism transformations, using symmetric local cross-correlation criterion as the similarity loss function. However, these models lack the constraints deploying temporal dependencies in the loss function to regulate the continuous motion in a sequence. More recently, Krebs et al. (2021) extended their probabilistic model to a spatio-temporal registration method for SAX cine CMR images. In this model, time dependencies are modelled using a temporal convolutional network (TCN) and a temporal dropout (TD) scheme to capture local dependencies over time. They performed motion simulation and motion transport by applying the recovered motion from one subject to another (Krebs et al., 2019b, 2021). Although the temporal dependencies were elegantly captured via a Gaussian process in the low dimensional latent space, no pixel-wise explicit probability distributions for the deformations were specified. The uncertainties in the estimated deformations remained unexplored and only cardiac cycle generation was demonstrated in terms of heart volumetric variations.

Another advantage of the probabilistic view over other learning-based methods is analytical uncertainty estimation. Clinicians benefit from this information in terms of the data analysis and confidence in the model for decision-making. Data related uncertainty (also referred to as aleatoric uncertainty) and uncertainty in the model parameters and structure (epistemic uncertainty) induce the predictive uncertainty (i.e., the confidence we have in a prediction) (Psaros et al., 2022). However, they are difficult to be assessed in a high-dimensional complex model, needing an uncertainty quantification approach proposed in Bayesian neural networks (Psaros et al., 2022; Wilson and Izmailov, 2020). One practical approach for approximate inference of the uncertainties is to execute stochastic forward pass when applying dropouts to weights (Gal and Ghahramani, 2015; Kendall et al., 2015; Kendall and Gal, 2017). However, this strategy increases inconsistent outputs (Kohl et al., 2018). By sampling from the learned velocity fields, propagating them through the diffeomorphic layers to calculate the deformation fields, and calculating the empirical diagonal covariance across samples, Dalca et al. (2018) describe an empirical method for estimating uncertainty for motion fields. However, the pixel-level deformation uncertainties are not explicitly modelled by this technique.

In summary, for motion tracking, unsupervised learning-based DIR approaches are preferred. The majority of the proposed methods recover deformations between independent pairs of images and, as a result, do not capture temporal dependencies. Only a few research provide probabilistic models to simulate motion for a particular application, and mainly do not include spatio-temporal modelling. Explicit modelling of deformation uncertainties is largely ignored in most of these studies. On the other hand, generating a realistic motion sequence given one frame is reported challenging in cardiac spatio-temporal motion modelling (Krebs et al., 2021).

In this study, to address these limitations, we propose a novel probabilistic spatio-temporal registration framework for cine cardiac MR imaging with two distinctive aims: (i) learning temporal motion fields by modelling the dependencies across the full range of cardiac frames, and (ii) generation of realistic CMR image sequences and their corresponding motion fields, which could be potentially applicable for motion simulation and data augmentation. Our main contributions are as follows:

- We propose an unsupervised statistical motion model, which uses a recurrent latent variable structure to infer probabilistic displacement fields in their original high dimensional spaces.

- We generate high temporal resolution image sequences given only one reference frame. To this end, the model learns to generate time dependent motion fields. This is accomplished through learning spatio-temporally-solved probability distributions for motion fields.
- The proposed probabilistic framework explicitly models the spatio-temporal uncertainty maps in the pixel level, which enables efficient analysis of the confidence in the derived heart motions without opting for empirical methods.

The remainder of this paper is organised as follows. The proposed model is described in Section 2. Section 3 presents the experimental analysis and results. We discussed the results in Section 4. Finally, Section 5 draws conclusions and signposts directions for future work.

2. Methodology

We leverage the generative nature of the variational autoencoder (VAE) in combination with recurrent neural network (RNN) (Chung et al., 2015) to introduce the Deformable Registration and Generative Network (DragNet) for predicting spatio-temporal DVFs in image sequences and generating synthetic datasets. Here, the purpose of the temporal registration algorithm is to compute the probability distribution of the DVFs per time step t ($p(\mathbf{D}_t)$), which spatially transforms image $\mathbf{I}_{t-1}$ (moving image) to the next image $\mathbf{I}_t$ (fixed image). Note that considering a reference moving image and finding deformations between that and all other images in the sequence is also possible in this framework (Section 3.7). Details of the method are described in the following including, the probabilistic DragNet, objective function, and the network architecture.

2.1. DragNet for probabilistic spatio-temporal registration

Given a temporal sequence of T images, $\{\mathbf{I}_0, \dots, \mathbf{I}_{T-1}\}$, acquired during a full cardiac cycle, consider expanding the joint distribution $p(\mathbf{I}_{\leq T-1}, \mathbf{z}_{\leq T-1}, \mathbf{D}_{\leq T-1})$ using the chain rule as

$$p(\mathbf{I}_{\leq T-1}, \mathbf{z}_{\leq T-1}, \mathbf{D}_{\leq T-1}) = \prod_{t=0}^{T-1} p(\mathbf{I}_t|\mathbf{I}_{<t}, \mathbf{z}_{\leq t}, \mathbf{D}_{\leq t}) p(\mathbf{z}_t|\mathbf{z}_{<t}, \mathbf{I}_{<t}, \mathbf{D}_{\leq t}) p(\mathbf{D}_t|\mathbf{D}_{<t}, \mathbf{z}_{<t}, \mathbf{I}_{<t}) \tag{1}$$

where the image at time t, $\mathbf{I}_t$, depends on a set of preceding images $\mathbf{I}_{<t}$, the latent variables $\mathbf{z}_{\leq t}$, and the displacement field maps $\mathbf{D}_{\leq t}$. The model learns the spatio-temporal dependencies between images via a recurrent Convolutional Long Short Term Memory (Conv-LSTM) (Shi et al., 2015). The dependencies among the preceding images, the latent variables, and the displacements field maps (i.e., $\mathbf{I}_{<t}, \mathbf{z}_{<t}$, and $\mathbf{D}_{<t}$) are captured through the hidden state variable $\mathbf{h}_{t-1}$ in the Conv-LSTM. Besides, at each time step, we condition the generative process only on the previous image $\mathbf{I}_{t-1}$ (moving image) and generate the current image $\mathbf{I}_t$ using the inferred distribution $\mathbf{D}_t$. Consequently, we assume the following factorisation replacing Eq. (1)

$$p(\mathbf{I}_{\leq T-1}, \mathbf{z}_{\leq T-1}, \mathbf{D}_{\leq T-1}) = \prod_{t=0}^{T-1} p(\mathbf{I}_t|\mathbf{I}_{t-1}, \mathbf{D}_t) p(\mathbf{z}_t|\mathbf{h}_{t-1}) p(\mathbf{D}_t|\mathbf{h}_{t-1}) \tag{2}$$

Here, we define $\mathbf{I}_{-1} = \mathbf{I}_{T-1}$, such that for $t = 0$ in Eq. (2), we have $p(\mathbf{I}_0|\mathbf{I}_{T-1}, \mathbf{D}_0)$, meaning that in the consecutive image registration framework, we use the last image in the sequence (i.e., $\mathbf{I}_{T-1}$) as the moving image to generate the image at the time 0. For the sake of simplicity, we assume that $\mathbf{I}_t$ is conditionally independent of $\mathbf{z}_{\leq t}$ but dependent on $\mathbf{D}_t$. Because of the $\mathbf{z}_t$ dependence on $\mathbf{h}_{t-1}$, we can indirectly relate $\mathbf{I}_t$ to $\mathbf{z}_t$ and also to the preceding frames by inferring the posterior distribution of the displacements from the latent variable $\mathbf{z}_t$, as described below. In Eq. (2), we condition the prior probability of the latent variables $\mathbf{z}_t$ on the $\mathbf{h}_{t-1}$, which is assumed to be a multivariate Gaussian distribution

$$p(\mathbf{z}_t|\mathbf{h}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \boldsymbol{\mu}_{\mathbf{z}_t,pi}, diag(\boldsymbol{\sigma}^2_{\mathbf{z}_t,pi})) \tag{3}$$

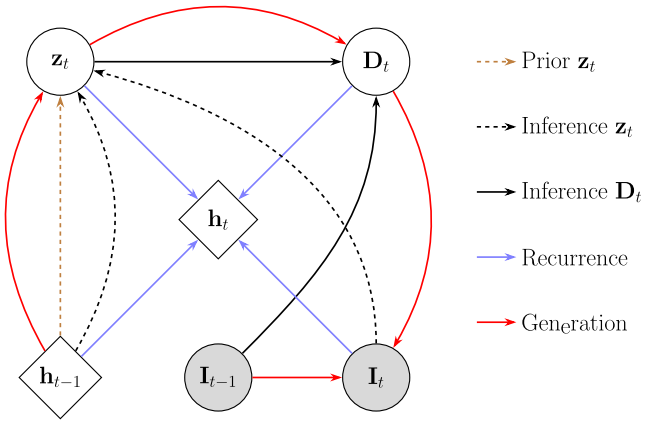

Fig. 1. Graphical representation of the model at time t, indicating the prior distribution of $\mathbf{z}_t$ based on Eq. (3) (in brown), inference of the approximate posterior probabilities of $\mathbf{z}_t$ and $\mathbf{D}_t$ using Eqs. (6) and (7), respectively, generating motion $\mathbf{D}_t$ and image $\mathbf{I}_t$ (in red), updating the hidden state variables of the RNN based on Eq. (9)(in blue).

where $\boldsymbol{\mu}_{\mathbf{z}_t,pi}$ and $\boldsymbol{\sigma}^2_{\mathbf{z}_t,pi}$ denote the mean and covariance of the distribution learned via a network. We also assume a Gaussian distribution for the prior of the displacement fields $\mathbf{D}_t$ with a zero mean and the identity covariance

$$p(\mathbf{D}_t|\mathbf{h}_{t-1}) = \mathcal{N}(\mathbf{D}_t; \mathbf{0}, \mathbf{1}) \tag{4}$$

We use the variational approach introduced by Kingma and Welling (2013) and assume that the approximate posterior of $\mathbf{z}_t$ and $\mathbf{D}_t$ can be factorised as

$$\begin{aligned} q(\mathbf{D}_{\leq T-1}, \mathbf{z}_{\leq T-1}|\mathbf{I}_{\leq T-1}) &= \prod_{t=0}^{T-1} q(\mathbf{z}_t|\mathbf{z}_{<t}, \mathbf{D}_{<t}, \mathbf{I}_{\leq t}) q(\mathbf{D}_t|\mathbf{D}_{<t}, \mathbf{z}_{\leq t}, \mathbf{I}_{\leq t}) \\ &= \prod_{t=0}^{T-1} q(\mathbf{z}_t|\mathbf{I}_t, \mathbf{h}_{t-1}) q(\mathbf{D}_t|\mathbf{I}_{t-1}, \mathbf{z}_t) \end{aligned} \tag{5}$$

where

$$q(\mathbf{z}_t|\mathbf{I}_t, \mathbf{h}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \boldsymbol{\mu}_{\mathbf{z}_t}, diag(\boldsymbol{\sigma}^2_{\mathbf{z}_t})) \tag{6}$$

Then, we take a sample, $\hat{\mathbf{z}}_t$, from $q(\mathbf{z}_t|\mathbf{I}_t, \mathbf{h}_{t-1})$ using the standard reparameterisation trick (Kingma and Welling, 2013). Features of this sample in concatenation with the feature sets from $\mathbf{I}_{t-1}$ form the input of a neural network that estimates $\boldsymbol{\mu}_{\mathbf{D}_t}$ and $\boldsymbol{\Sigma}_{\mathbf{D}_t}$, the parameters of the multivariate Gaussian posterior distribution of $\mathbf{D}_t$ defined as

$$q(\mathbf{D}_t|\mathbf{I}_{t-1}, \hat{\mathbf{z}}_t) = \mathcal{N}(\mathbf{D}_t; \boldsymbol{\mu}_{\mathbf{D}_t}, \boldsymbol{\Sigma}_{\mathbf{D}_t}) \tag{7}$$

Two inference networks implement the Eqs. (6) and (7), predicting the parameters of the Gaussian distributions from their corresponding inputs. For image generation at time t, we first draw a sample displacement field $\hat{\mathbf{D}}_t \sim q(\mathbf{D}_t|\mathbf{I}_{t-1}, \hat{\mathbf{z}}_t)$ and then obtain $\mathbf{I'}_t$ image by warping the $\mathbf{I}_{t-1}$ via the displacement fields $\hat{\mathbf{D}}_t$ using a spatial transformer network (STN) (Jaderberg et al., 2015):

$$\mathbf{I'}_t = \mathcal{T}(\mathbf{I}_{t-1}, \hat{\mathbf{D}}_t) \tag{8}$$

where $\mathcal{T}$ is the spatial transformation (resampling) module.

Finally, the Conv-LSTM module updates its hidden state variables using the recurrence equation

$$\mathbf{h}_t = ConvLSTM\left(\mathbf{I}_t, \mathbf{z}_t, \mathbf{D}_t, \mathbf{h}_{t-1}\right) \tag{9}$$

Fig. 1 shows the graphical representation of our models for implementing the prior, variational inference, generative, and the recurrence path at time t.

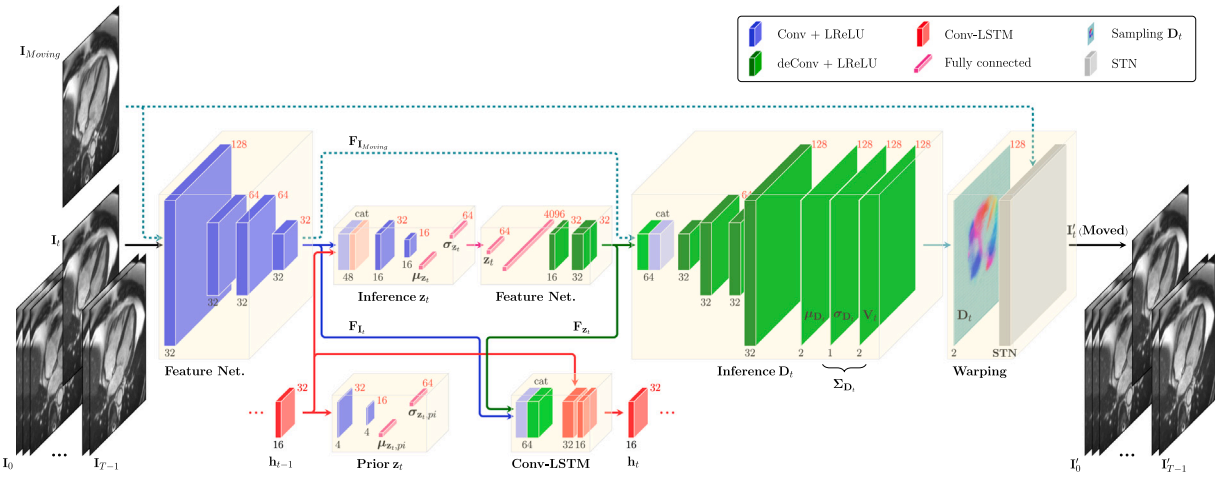

Fig. 2. The architecture of the proposed DragNet is illustrated. The model comprises of five main modules: neural networks to compute the parameters of the prior and the posterior distributions of the latent variables ($\mathbf{z}_t$), posterior distributions of the displacement field ($\mathbf{D}_t$), the deterministic recurrent parameters ($\mathbf{h}_t$) via Conv-LSTM layer, and a spatial transformer network (STN) layer that warps the moving image from the previous frame (or a specific reference frame) to the fixed image at time t. The CMR images were reproduced with a permission of UK Biobank©.

2.2. Objective function

The generative model $p(\mathbf{I}_{\leq T-1}, \mathbf{z}_{\leq T-1}, \mathbf{D}_{\leq T-1})$, Conv-LSTM network, and the inference model $q(\mathbf{D}_{\leq T-1}, \mathbf{z}_{\leq T-1}|\mathbf{I}_{\leq T-1})$ are trained jointly by maximising a variational Evidence Lower Bound (ELBO) (Kingma and Welling, 2013) with respect to their parameters using stochastic gradient methods. Moreover, to control the smoothness of the predicted motion fields, we use a smoothness loss used in previous learning-based methods by Yu et al. (2016) and Balakrishnan et al. (2019). Therefore, The overall loss function is given by

$$\mathcal{L} = -\mathcal{L}_{ELBO} + \lambda \mathcal{L}_{Smoothness} \tag{10}$$

where λ is a regularisation parameter that balances the registration accuracy and the smoothness of the predicted displacements. In our experiments, we have set $\lambda = 0.03$. $\mathcal{L}_{ELBO}$ is defined as

$$\mathcal{L}_{ELBO} = \mathbb{E}_{q(\mathbf{D}_{\leq T-1}, \mathbf{z}_{\leq T-1}|\mathbf{I}_{\leq T-1})} \log \frac{p(\mathbf{I}_{\leq T-1}, \mathbf{z}_{\leq T-1}, \mathbf{D}_{\leq T-1})}{q(\mathbf{D}_{\leq T-1}, \mathbf{z}_{\leq T-1}|\mathbf{I}_{\leq T-1})} \tag{11}$$

Using Eqs. (2) and (5), the ELBO term can be written as

$$\mathcal{L}_{ELBO} = \mathbb{E}_{\prod_{t=0}^{T-1} q(\mathbf{z}_t|\mathbf{I}_t, \mathbf{h}_{t-1}) q(\mathbf{D}_t|\mathbf{I}_{t-1}, \mathbf{z}_t)} \left[\sum_{t=0}^{T-1} \log p(\mathbf{I}_t|\mathbf{I}_{t-1}, \mathbf{D}_t) + \log \frac{p(\mathbf{z}_t|\mathbf{h}_{t-1})}{q(\mathbf{z}_t|\mathbf{I}_t, \mathbf{h}_{t-1})} + \log \frac{p(\mathbf{D}_t|\mathbf{h}_{t-1})}{q(\mathbf{D}_t|\mathbf{I}_{t-1}, \mathbf{z}_t)} \right] \tag{12}$$

which can be decomposed into three terms as follows

$$\mathcal{L}_{ELBO} = \mathcal{L}_{Sim} + \mathcal{L}_{\mathbf{z}} + \mathcal{L}_{\mathbf{D}} \tag{13}$$

where $\mathcal{L}_{Sim}$ controls the similarity between the warped image obtained from Eq. (8) and the fixed image at each time step (i.e., $\mathbf{I}_t$). On the other hand, $\mathcal{L}_{\mathbf{z}}$ and $\mathcal{L}_{\mathbf{D}}$ constrain the probability distribution of the latent variables $\mathbf{z}_t$ and $\mathbf{D}_t$. We derive the $\mathcal{L}_{Sim}$ as

$$\mathcal{L}_{Sim} \simeq \sum_{t=0}^{T-1} \frac{1}{L} \sum_{l=1}^{L} \mathbb{E}_{q(\mathbf{D}_t|\mathbf{I}_{t-1}, \mathbf{z}_t^{(l)})} \log p(\mathbf{I}_t|\mathbf{I}_{t-1}, \mathbf{D}_t) \tag{14}$$

where the expectation over $q(\mathbf{z}_t|\mathbf{I}_t, \mathbf{h}_{t-1})$ is taken empirically using L Monte Carlo samples. In a similar manner, we have:

$$\mathbb{E}_{q(\mathbf{D}_t|\mathbf{I}_{t-1}, \mathbf{z}_t^{(l)})} \log p(\mathbf{I}_t|\mathbf{I}_{t-1}, \mathbf{D}_t) \simeq \frac{1}{K} \sum_{k=1}^{K} \log p\Big(\mathbf{I}_t|\mathbf{I}_{t-1}, \mathbf{D}_t^{(k)}(\mathbf{z}_t^{(l)})\Big) \tag{15}$$

where $\mathbf{D}_t^{(k)}(\mathbf{z}_t^{(l)})$ indicates the kth DVF sample as a function of $\mathbf{z}_t^{(l)} \sim q(\mathbf{z}_t|\mathbf{I}_t, \mathbf{h}_{t-1})$. Hence Eq. (14) can be simplified as

$$\begin{aligned} \mathcal{L}_{Sim} &\simeq \frac{1}{LK} \sum_{t=0}^{T-1} \sum_{l=1}^{L} \sum_{k=1}^{K} \log p\Big(\mathbf{I}_t|\mathbf{I}_{t-1}, \mathbf{D}_t^{(k)}(\mathbf{z}_t^{(l)})\Big) \\ &\simeq \frac{1}{LK} \sum_{t=0}^{T-1} \sum_{l=1}^{L} \sum_{k=1}^{K} \|\mathbf{I}_t - \mathcal{T}(\mathbf{I}_{t-1}, \mathbf{D}_t^{(k)}(\mathbf{z}_t^{(l)}))\|^2 \end{aligned} \tag{16}$$

The gradient of this approximation can be backpropagated with the reparameterisation trick (Kingma and Welling, 2013).

The second term in the ELBO (Eq. (12)) denotes the Kullback–Leibler divergence (KL divergence) (Hershey and Olsen, 2007) between the approximate posterior $q(\mathbf{z}_t|\mathbf{I}_t, \mathbf{h}_{t-1})$ and the prior distribution $p(\mathbf{z}_t|\mathbf{h}_{t-1})$ and is given by

$$\mathcal{L}_{\mathbf{z}} = -\sum_{t=0}^{T-1} \mathcal{D}_{\mathcal{KL}}\Big(q(\mathbf{z}_t|\mathbf{I}_t, \mathbf{h}_{t-1}) \parallel p(\mathbf{z}_t|\mathbf{h}_{t-1})\Big) \tag{17}$$

which is computed analytically. Similarly, by drawing L Monte Carlo samples from $q(\mathbf{z}_t|\mathbf{I}_t, \mathbf{h}_{t-1})$, the KL divergence between the estimated posterior and prior of $\mathbf{D}_t$ can be computed using

$$\mathcal{L}_{\mathbf{D}} = -\frac{1}{L} \sum_{t=0}^{T-1} \sum_{l=1}^{L} \mathcal{D}_{\mathcal{KL}}\Big(q(\mathbf{D}_t|\mathbf{I}_{t-1}, \mathbf{z}_t^{(l)}) \parallel p(\mathbf{D}_t|\mathbf{h}_{t-1})\Big) \tag{18}$$

The detailed computations of $\mathcal{L}_{\mathbf{z}}$ and $\mathcal{L}_{\mathbf{D}}$ are given in Appendices A and B.

$\mathcal{L}_{Smoothness}$ in Eq. (10), acts as a diffusion regulariser and encourages a smooth displacement field $\mathbf{D}_t$ by taking spatial gradients of displacements over the entire cardiac cycle

$$\mathcal{L}_{Smoothness} = \sum_{t=0}^{T-1} \|\nabla \mathbf{D}_t\|^2 \tag{19}$$

We use stochastic gradient descent based methods to optimise the total loss function (Eq. (10)) and learn the parameters of the network.

2.3. Network architecture and implementation

As Fig. 2 illustrates, the architecture of the model consists of five different components:

- *Module one (Prior of $\mathbf{z}_t$)*: This module computes the parameters of the prior distribution of the latent variable $\mathbf{z}_t$ using Eq. (3). We define $\mathbf{h}_{t-1} \in \mathbb{R}^{C \times M \times N}$, where C and $M \times N$ indicate the channel and spatial dimensions, respectively. In our implementation, we

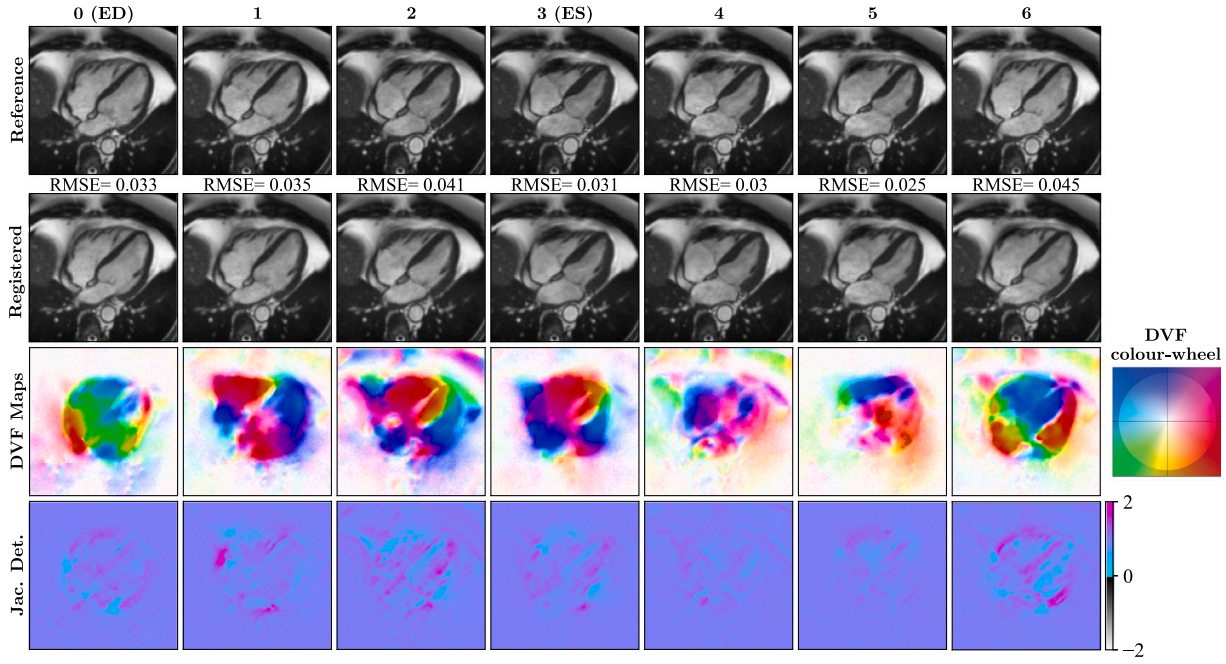

Fig. 3. Spatio-temporal registration between consecutive frames in an LAX CMR image sequence consisting of seven frames. The registered image at the first time point is obtained by warping the image at time point 7 to the ED frame. The registered images show overall good compatibility with the reference images. The corresponding colour-coded DVF maps showing the direction of displacements are smooth and mostly have positive Jacobian determinants. The reference CMR images were reproduced with a permission of UK Biobank©.

set C, M, and N as 16, 32, and 32, respectively. We pass $\mathbf{h}_{t-1}$ through the convolutional layers, then reshape the output to be compatible for feeding through fully connected (FC) layers that compute $\boldsymbol{\mu}_{\mathbf{z}_t,pi}$ and $\boldsymbol{\sigma}^2_{\mathbf{z}_t,pi} \in \mathbb{R}^d$. In our experiments, we have set the latent space dimension as $d = 64$.

- *Module two (Inference of $\mathbf{z}_t$)*: Here, we compute the parameters of the posterior distribution of the latent variable $\mathbf{z}_t$ according to Eq. (6), where a CNN generated feature map of $\mathbf{I}_t \in \mathbb{R}^{1\times H\times W}$ (i.e., $\mathbf{F}_{\mathbf{I}t} \in \mathbb{R}^{32\times 32\times 32}$ in Fig. 2) is concatenated with $\mathbf{h}_{t-1}$ and used as input. We sample $\mathbf{z}_t \in \mathbb{R}^d$ and apply it as input to the next module.
- *Module three (Inference of $\mathbf{D}_t$)*: In this module, we compute the parameters of the posterior distribution for the displacement field ($\mathbf{D}_t$) based on Eq. (7). The sampled $\mathbf{z}_t$ from module 2 is applied as an input to a fully connected layer and then reshaped to a 3D tensor ($\mathbf{F}_{\mathbf{z}t}$ in Fig. 2) to be concatenated with $\mathbf{F}_{\mathbf{I}Moving}$ (i.e., the feature map of the moving image). Finally, after passing the latter through the deconvolution layers in Fig. 2, this module generates the tensors of $\boldsymbol{\mu}_{\mathbf{D}_t}, \mathbf{V}_t \in \mathbb{R}^{2\times H\times W}$, and $\boldsymbol{\sigma}^2_{\mathbf{D}_t} \in \mathbb{R}^{1\times H\times W}$ that determine the mean of the $\mathbf{D}_t$'s posterior distribution as its covariance matrix is given using

$$\boldsymbol{\Sigma}_{\mathbf{D}_t} = \boldsymbol{\sigma}^2_{\mathbf{D}_t}\mathbf{1} + \mathbf{V}_t\mathbf{V}_t^{\intercal} \tag{20}$$

  where $\mathbf{1} \in \mathbb{R}^{2\times 2\times H\times W}$ represents an identity matrix.
- *Module four (Warping)*: This module consists of a spatial transformer network layer that, given a displacement field $\mathbf{D}_t$ and moving image, resamples the latter at the pixel locations specified by $\mathbf{D}_t$, warping the image.
- *Module five*: Here, a Conv-LSTM network is implemented to model the temporal dependencies of the latent and deformation variables. As shown in Fig. 2, at each time step, this module updates the hidden state variables based on the features of latent variables and the fixed image at time t. The updated hidden state plays a critical role in the estimation of the distribution of $\mathbf{z}_{t+1}$ and $\mathbf{D}_{t+1}$.

The whole framework is implemented using Python and PyTorch. The Adam optimiser (Kingma and Ba, 2014) is used for optimising the loss function, with the learning rate of 0.001.

3. Experiments and results

3.1. Utilised dataset

In this study, a dataset from UK Biobank (UKB) LAX cine CMR images (4-chamber view) comprising 4620 subjects (mean age 58.7, 52.5% female) (Petersen et al., 2015) is analysed. The balanced steady-state free precession (bSSFP) cine acquisition was used. Each cardiac cycle imaged at 50 frames with the matrix size of 208 × 187 and the in-plane image resolution of 1.8×1.8 mm^2. More details on the image acquisition protocol can be found in Petersen et al. (2015). Ground-truth contours of the left and right atria (LA/RA) at the end-diastolic (ED) and end-systolic (ES) frames, which were annotated by clinical experts, are also used to evaluate the model performance. Each sequence starts from the ED frame. All images were cropped to the equal size of 128 × 128 pixels to cover the whole heart avoiding redundant areas. Moreover, the image intensities were normalised to the range of $[0, 1]$ before feeding to the network. We have trained the model using 4000 image sequences, which were selected randomly from the population. The rest of the 620 subjects were used for the evaluation. Each batch of training consists of 10 cardiac sequences.

3.2. Baseline registration methods and evaluation metrics

We compare our model, in terms of registration accuracy and spatio-temporal deformation regularity, with five state-of-the-art deformable image registration methods: Demons (Vercauteren et al., 2008, 2009), ANTs SyN (Avants et al., 2008, 2011), NiftyReg (Modat et al., 2010), a learning-based model (VoxelMorph) (Balakrishnan et al., 2019), and temporal B-spline algorithm in Elastix (2D + t (3D)-Elastix) (Metz et al., 2011). The diffeomorphic Demons approach is implemented based on the original algorithm in Vercauteren et al. (2008) and Vercauteren et al. (2009) and optimised on the CMR images in three-level resolution with the number of iterations of 250. For the ANTs SyN, we use the Symmetric Normalisation (SyN) implementation in the publicly available Advanced Normalisation Tools (ANTs) software package (Avants et al., 2011), with a mutual information similarity

Table 1
Registration results (mean ± std) between consecutive frames on test sequences for different methods in terms of RMSE, the number of pixels with negative Jacobian determinant, Dice scores at ED and ES when deforming LA+RA masks sequentially from ED towards ES and vice versa, and the temporal gradients of the displacement fields. **Bold** values indicate the best results among different techniques in each column.

Method	RMSE	$\#\lvert J_{\mathbf{D}}\rvert \leq 0$	Dice (%) at ED	Dice (%) at ES	Temporal Grad.
Demons	0.032 ± 0.01	3311	76.1 ± 6.2	85.5 ± 5.5	0.70 ± 0.22
ANTs SyN	0.037 ± 0.01	4060	79.4 ± 5.7	**90.7 ± 2.9**	0.53 ± 0.09
NiftyReg	0.033 ± 0.01	2484	80.9 ± 5.3	88.5 ± 2.8	0.65 ± 0.14
VoxelMorph	**0.030 ± 0.01**	3409	80.8 ± 4.9	90.6 ± 2.9	0.46 ± 0.09
2D+t-Elastix	0.036 ± 0.01	6369	80.4 ± 6.7	86.8 ± 5.8	**0.32 ± 0.07**
DragNet	0.032 ± 0.01	**2118**	**81.1 ± 4.6**	89.4 ± 3.0	0.33 ± 0.08

measure. The gradient step size is set to 0.6, and smoothing for update fields is 5 at three scales with 60 iterations each. We use the NiftyReg package based on the Free-Form Deformation algorithm for non-rigid registration and utilise the CPU version, which is publicly available by the authors Modat et al. (2010). Specifically, we used the Localised Normalised Cross Correlation (LNCC) objective function, grid spacing of 5, and the number of iterations of 300. The temporal Elastix registration method is implemented based on the publicly available python package and the original paper (Metz et al., 2011). We tuned it for the given cine cardiac MRI by setting grid spacing of 8 and number of iterations of 256. We implement VoxelMorph method with cross-correlation similarity measure, based on a publicly available package and trained the model using 4662 image pairs from the LAX CMR dataset for 40 epochs with the learning rate of 0.001.

To measure the registration accuracy, we use the root mean square error (RMSE) of intensities between the fixed and the warped images. We also evaluate the diffeomorphic (invertibility) property of the registration algorithm by computing the Jacobian matrix $J_{\mathbf{D}_{tk}} = \nabla \mathbf{D}_{tk} \in \mathbb{R}^{2\times 2}$ that captures the local properties of deformation around pixel k. The local deformation is diffeomorphic, only at locations for which $|J_{\mathbf{D}_{tk}}| > 0$ (Ashburner, 2007). To assess the diffeomorphic property, we count the number of pixels where $|J_{\mathbf{D}_{tk}}| \leq 0$. The estimated temporal deformations are assessed by deforming the LA/RA masks at the ED phase towards the ES phase and vice versa. Then, the Dice similarity coefficient (Dice, 1945) is utilised to assess the motion fields deforming the anatomical structures of LA and RA over time. A Dice score of one indicates the maximum anatomical match, whereas a score of zero shows no overlap. A statistical paired t-test is applied to determine if there is a significant difference between the groups of measured metrics.

To assess the quality of the generated images, we use structural similarity index measure (SSIM) (Wang et al., 2004) between the original and the generated sequences. To investigate how the generated motion sequences resemble the actual motion of a typical heart (motion quality assessment), we measure variations of the area of the simulated LA/RA over time when compared to the original variations.

3.3. Registration of consecutive frames

In this experiment, we evaluate the model performance on estimating the DVFs between successive CMR images within sequences consisting of seven frames showing clear distinctive appearances. The original sequences were downsampled in time, such that the first frame represents ED and the fourth frame corresponds to ES. Fig. 3 represents an example set of results of spatio-temporal registration using the proposed framework. The results show that there is generally a good match between the reference and the registered images. The DVF maps are smooth, leading to positive Jacobian determinant values. The colour wheel shows the magnitude and direction of motion fields, where the angle with the x-axis indicates the motion direction, and the colour intensity expresses the magnitude of the displacement.

Table 1 shows the performance of the proposed model (DragNet) compared with five other state-of-the-art techniques on the test sequences in terms of: registration accuracy, the number of locations with non-positive Jacobian determinants, Dice scores for the anatomical structures, and the temporal gradient of the displacement fields. Detailed box plots of the results are represented in Fig. 4. The results indicate that DragNet performs temporal registration within RMSE ranges comparable to those obtained with Demons, NiftyReg, and VoxelMorph, but significantly better than ANTs, and 3D-Elastix ($p < 0.05$). In addition, DragNet presents fewer negative Jacobian determinant locations than other algorithms in the consecutive frame registrations (Table 1 and Fig. 4). In terms of Dice scores, when deforming anatomical structures sequentially from ED towards ES and vice versa, the proposed model significantly outperforms the Demons and 3D-Elastix algorithms at ES ($p < 0.05$) but is on par with VoxelMorph and NiftyReg ($p = 0.31$ and $p = 0.64$, respectively) (Table 1). Fig. 5 shows an example set of results for deforming LA + RA contours from ED towards ES phases in a sequence for all six methods.

The comparable results in terms of Dice and RMSE and lower temporal gradients of the displacements indicate that DragNet has the ability to significantly improve the temporal regularity of the deformation fields over the learning-based VoxelMorph model and all the other conventional pair-wise image registration algorithms ($p < 0.001$) (Table 1 and Fig. 4). This feature is justified by the limitation of pair-wise image registration methods on estimating the motion as a sequence of independent deformation fields. In contrast, DragNet considers the temporal dependencies when computing the deformation fields between pairs of images (via hidden states of the RNN). Compared to the temporal 2D + t-Elastix algorithm, DragNet presents similar temporal gradients of the displacements on average; however, it generally shows improved performance in other registration metrics (Table 1 and Fig. 4).

3.4. Uncertainty assessment

DragNet enables estimation of the spatio-temporal uncertainties in the DVFs through computing $\boldsymbol{\Sigma}_{\mathbf{D}_t}$ from Eq. (20):

$$H(\mathbf{D}_t) \approx \mathbb{E}[-\log q(\mathbf{D}_t|\mathbf{I}_{t-1}, \mathbf{z}_t)] = \frac{1}{2}\log\Big((2\pi)^2|\boldsymbol{\Sigma}_{\mathbf{D}_t}|\Big) \tag{21}$$

Fig. 6 shows the motion uncertainty maps in three different cardiac sequences.[2] The uncertainty values are lower in the anatomical boundaries, suggesting that the deformations are generally driven locally by high contrast areas such as boundaries of the heart. To assess the local variations, we have computed the average and the standard deviation of the uncertainties inside the anatomical structures of LA/RA and around the boundaries at ED and ES frames among the test dataset. The results reported in Table 2 and Fig. 7 indicate that the uncertainties around the boundaries are significantly lower than those within the hearts ($p < 0.001$). The homogeneous areas that do not move significantly (mainly outside the heart) show similar uncertainty values over time (see the Supplementary Materials). The local covariance matrix $\boldsymbol{\Sigma}_{\mathbf{D}_t}$ values are larger in these regions due to the lack of local imaging features, resulting in higher uncertainty values there according to Eq. (21). Similarly, due to the absence of regional characteristics

[2] A GIF file showing the animated uncertainty maps is also presented in the Supplementary Materials (Uncertainty_animation).

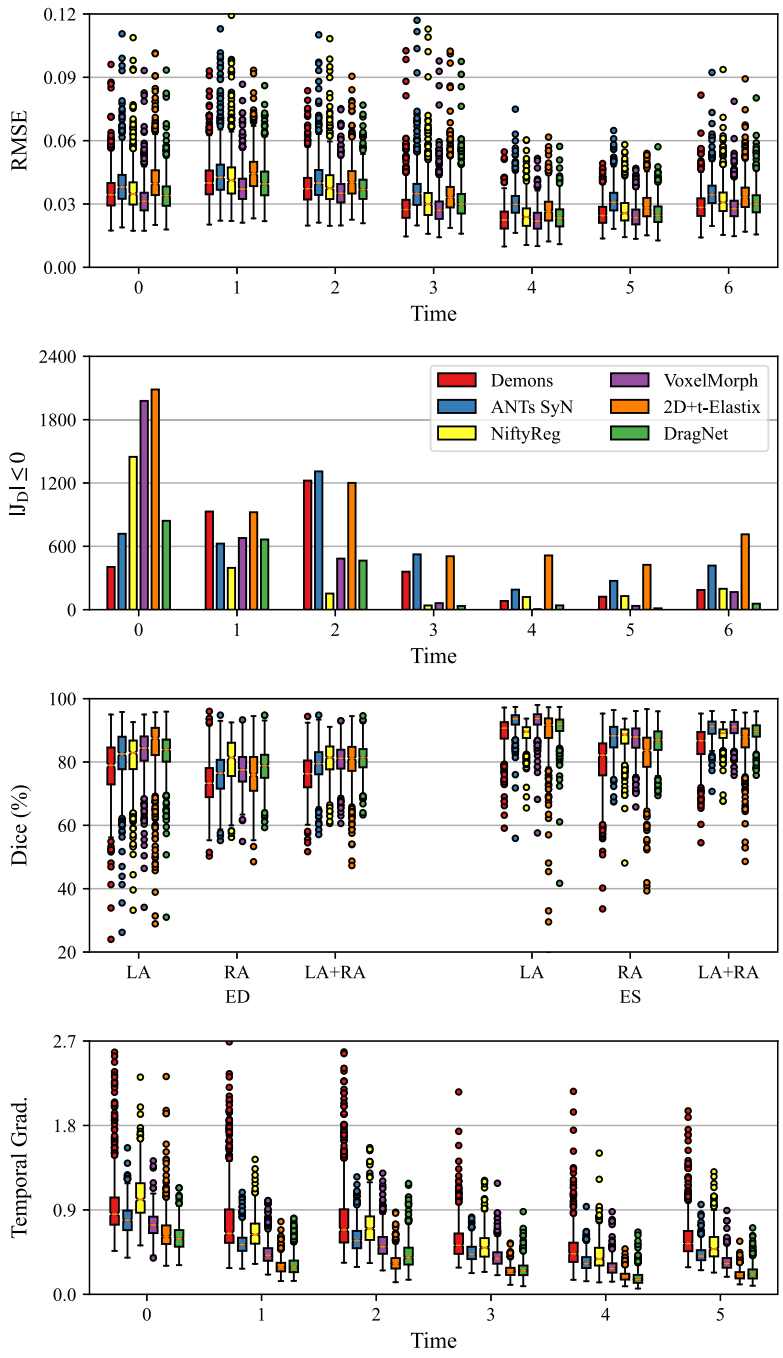

Fig. 4. The results of spatio-temporal registration between successive frames of test sequences showing RMSE, the number of non-positive Jacobian determinant locations, the Dice scores for anatomical structures, and the temporal gradients of the displacement fields. DragNet presents the registration results comparable to or even better than some state-of-the-art methods, while on par with temporal Elastix considerably improves temporal gradients.

in the blood pool motions, one can observe overall lower confidence in registrations within those areas compared to the boundaries. The uncertainty range for the example subjects shown in Fig. 6 is [−1.4, −0.8], which is the outcome of the logarithm function applied to the covariance matrix values ranging from 6e−7 to 0.08, considering both x and y directions.

3.5. Generation of synthetic motion sequences

As mentioned, in addition to registration, our proposed DragNet model can generate synthetic cine LAX CMR images given only one reference frame $\mathbf{I}_0$. To demonstrate this, we utilised the ED frame as the reference image to generate a full sequence accordingly. Fig. 8 shows two samples of generated cine CMR image sequences together with their corresponding DVF and Jacobian determinant maps. The original sequence is also shown for comparison. The results show that the

Table 2
Local uncertainty measurement (mean ± std) of motion fields inside the LA/RA area and around the borders. The results show a significant difference in the uncertainty values between the two regions (statistical significance p-value < 0.001) at both ED and ES phases.

Region	ED	ES
Inside LA/RA	−1.084 ± 0.021	−1.062 ± 0.014
Boundaries	−1.168 ± 0.026	−1.146 ± 0.021

Fig. 5. Tracking LA + RA contours over time from ED frame towards ES using deformations estimated by Demons, ANTs SyN, NiftyReg, VoxelMorph, 3D-Elastix, and the proposed model, DragNet. Comparison with the ground-truth (GT) contours at ES (last column) in terms of Dice scores indicates that DragNet after NiftyReg outperforms the other methods in the temporal evolution of LA/RA structures. The CMR images were reproduced with a permission of UK Biobank©.

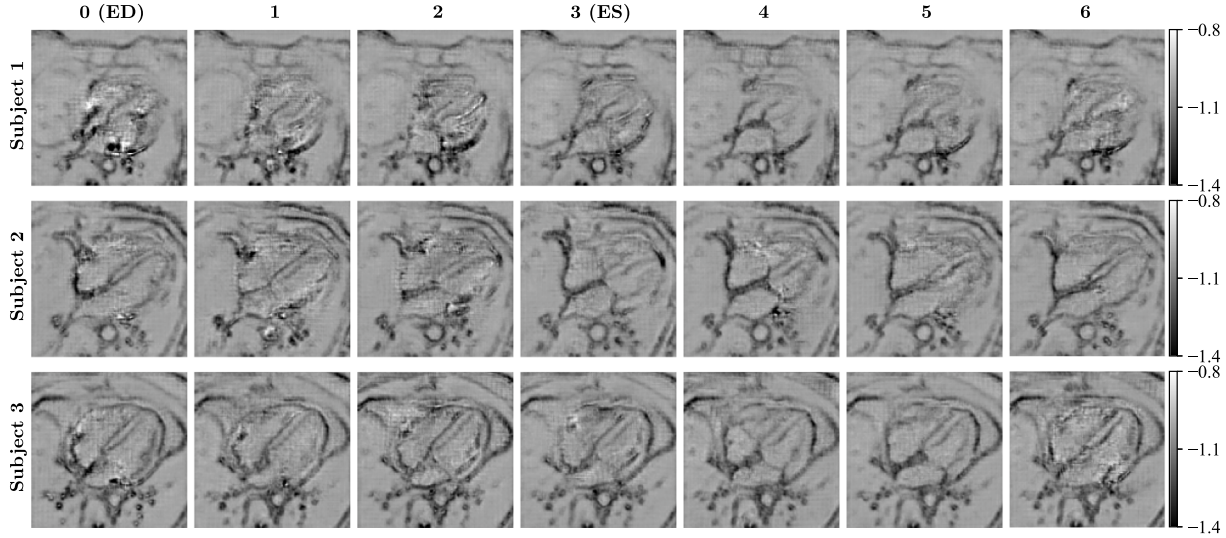

Fig. 6. Spatio-temporal DVF uncertainty maps for three different subjects, indicating lower uncertainties in the heart boundaries.

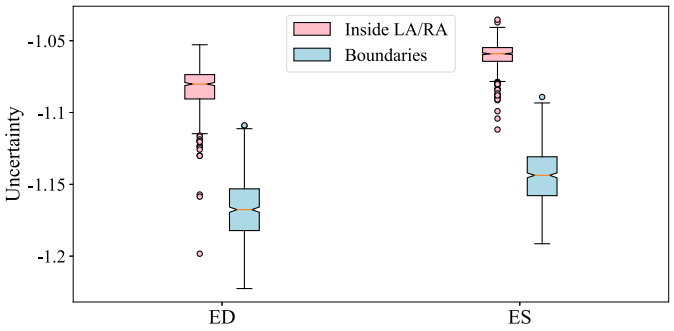

Fig. 7. The distribution of average DVF uncertainty values inside LA/RA and around the borders across the test population.

generated motion and image sequences are different in two generated instances and non-identical to the original one. Moreover, the generated motion sequences are also diffeomorphic with positive Jacobian determinant maps. Samples of the generated sequences with different temporal resolution are provided in the Supplementary Materials.

In contrary to the conventional generative models such as VAE that assume a Gaussian distribution with zero mean and identity covariance for the prior distribution of the latent variables $\mathbf{z}_t$, DragNet learns the prior distribution at each time from the history information captured using $\mathbf{h}_{t-1}$. This temporal modelling allows for generating realistic cardiac motion sequences in a cycle. To show this effect, we perform an ablation study by discarding the LSTM module and assume a unique Gaussian distribution for the prior distribution of $\mathbf{z}_t$ similar to the regular variational autoencoders. The result of a generated sequence is presented in Fig. 9, along with a sample sequence generated by considering $\mathbf{h}_t$. The generated sequence without considering $\mathbf{h}_t$ does not reflect the actual evolution of the anatomical structures, which should occur similarly to the original sequence in that at ED, the heart should show contracted atria and dilated ventricles, and as it moves towards ES, these conditions should be reversed. In the shown generation example without $\mathbf{h}_t$, one observes a random generation of cardiac MRI at each time point that does not follow a typical motion. However, when we consider the temporal modelling via LSTM hidden states, the generated cardiac sequence resembles a realistic motion.

3.6. Image and motion quality assessment of the generated sequences

In this experiment, we generate 100 synthetic heart motion samples based on single or more frames observed from each subject of the test dataset, generating a total of 62 000 synthetic cardiac images. We evaluate the quality of the acquired images using the SSIM (Wang et al., 2004), a perception-based model that considers image luminance (brightness), contrast (texture variations), and structural information when comparing the real to synthetic images. Fig. 10 represents the SSIM values between the test and the corresponding synthetic sequences obtained from DragNets delivering seven (Fig. 10a) and fourteen frames (Fig. 10b) per time point. Having observed image $\mathbf{I}_0$, generates $\mathbf{I}_1$ and the rest of the sequence including the cardiac frame at time point zero. We also investigate the impact of observing more frames on the quality of the synthesised images.

As shown in Fig. 10, the generated images have SSIM values mainly in the range of 70%–90% indicating an acceptable quality in terms of luminance, contrast, and presentation of structures when compared with the original images in the test dataset. Feeding more than one frame to the generation process increases the SSIM values in the starting time points. However, it tends to decrease to values even lower than those when only $\mathbf{I}_0$ were observed. The reason is that the model learns to capture the time-specific characteristics when updating the hidden state variables, such that starting only from the frame $\mathbf{I}_0$, it converges to a better local minimum with a lower value of the loss function. If more frames are provided to the model the generation starts from more detailed information in the initial frames causing large SSIM values. However, the converged local minimum is more constrained by the information from the initial frames and eventually becomes more suboptimal. The effect is pronounced when generating longer sequences (Fig. 10b).

To assess the quality of the generated motion sequences, we also investigate the variations of LA + RA area across the whole cardiac cycle and compare those with the motion in the test dataset. To this end, we warp the ground-truth LA + RA masks from ED phases towards the end of the cycle using 62K generated motion sequences discussed before. The box plots presenting the normalised area of the generated structures, along with the corresponding variations in the reference test dataset in red are shown in Fig. 11 (see the top row). The results are presented for seven-frame and fourteen-frame motion models. As shown, the synthetic motion sequences result in a pattern similar to the original test samples within an acceptable range of variations, indicating the diversity of the generation. The plots in the bottom row of Fig. 11 show the LA + RA areas computed from 100 synthetic sequences generated using a limited number of frames of a test sample. One can see that providing the model with more than one observed frame results in similar area values to those of the test sample in the initial frames at the cost of generating less divergent frames. Furthermore when the model observes more initial frames, a larger deviation from the real cardiac motion is obtained towards the end of the sequence. The improvement in the SSIM criterion alongside the larger range of variations in LA + RA motion in the synthetic sequences suggests that DragNet can generate realistic high temporal resolution cardiac motion sequences from a single frame.

3.7. Registration from the ED phase to the other phases

In this section, we evaluate the performance of the DragNet when the ED frame is registered to other cardiac frames and therefore larger displacement fields are generated. This experiment is performed on the seven-frame CMR sequences discussed in Section 3.3 and six DVF maps are computed for each sequence. Fig. 12 shows the results of different registration methods in a sample sequence. The registered ED images using DragNet are generally in good agreement with the target frames, where the RMSE values are comparable with the learning-based VoxelMorph and other state-of-the-art image registration techniques while presenting spatio-temporally smoother DVFs. Table 3 shows the quantitative results obtained using different registration methods on the test dataset. As shown DragNet achieves RMSE errors similar to those, shows less folding compared to the VoxelMorph, and outperforms all methods in terms of preserving more temporal DVF continuities ($p < 0.05$). The box plots in Fig. 13 compare the performance of the methods for all time points. The time point 2 in this figure shows the registration results from ED to ES. As it can be seen, the RMSE and the number of negative Jacobian determinant pixels initially increase slightly and then decrease towards the end of the sequence. This is because the largest displacements occur when registering the ED to ES phases.

Here, DragNet infers the large displacements between the ED and ES based on the RNN hidden state variables, which convey the information from the smaller previous displacements. Therefore, the model learns to build up large motion fields recursively by refining the previously observed smaller displacement, a difficulty that other algorithms address by taking a multi-scale registration approach.

4. Discussion

In this paper, we derive probabilistic motion fields for cardiac motion by formulating an unsupervised motion model based on the recurrent variational Bayes. Explicit modelling of the mean and variance

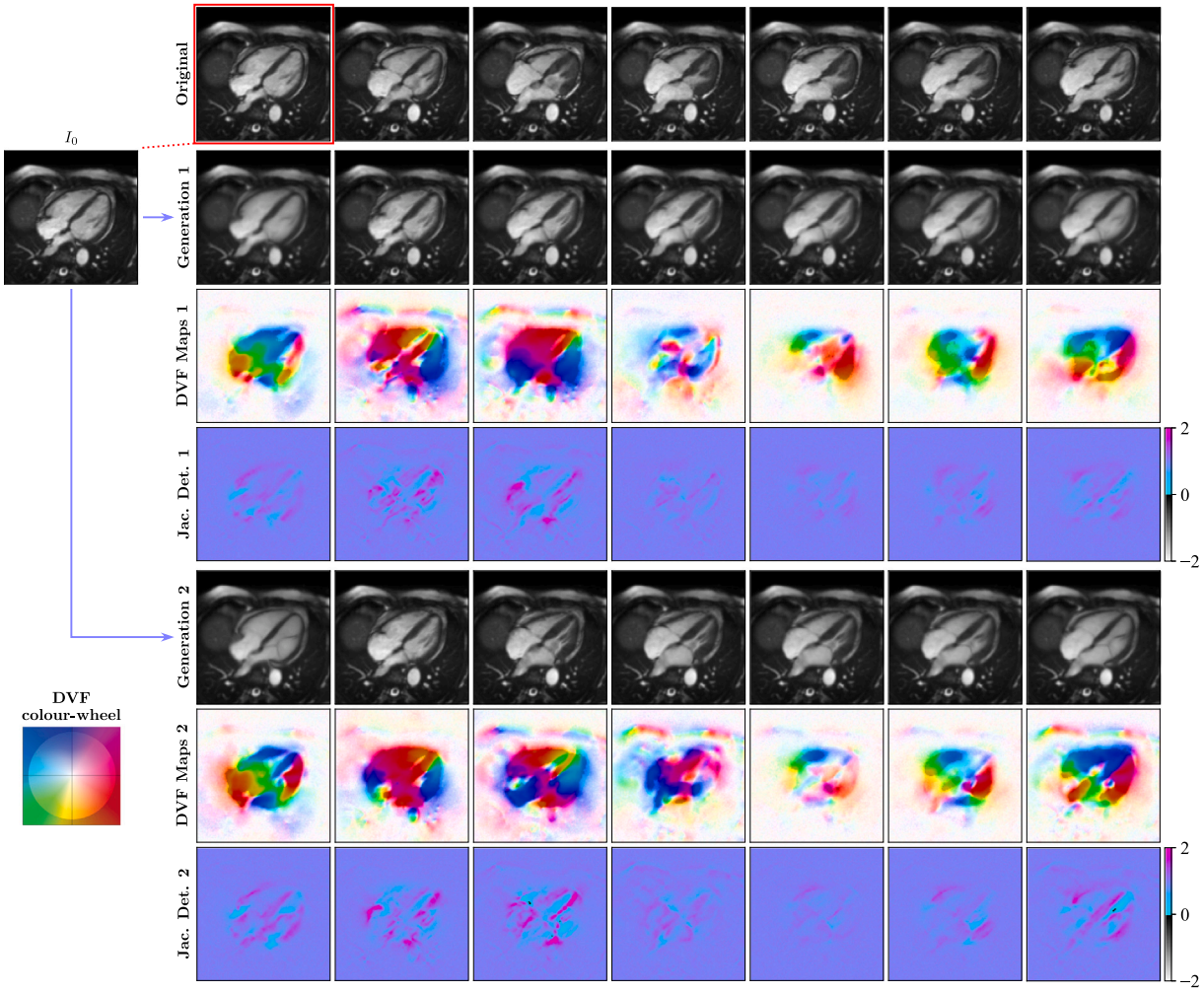

Fig. 8. Generation of two sample sequences of LAX CMR images along with the corresponding motion sequences and the Jacobian determinant maps, both generated from a reference image $\mathbf{I}_0$ (ED frame). The generated images are different from the original sequence, and the generated DVF sequences show invertibility by positive Jacobian determinants. The original CMR images were reproduced with a permission of UK Biobank©.

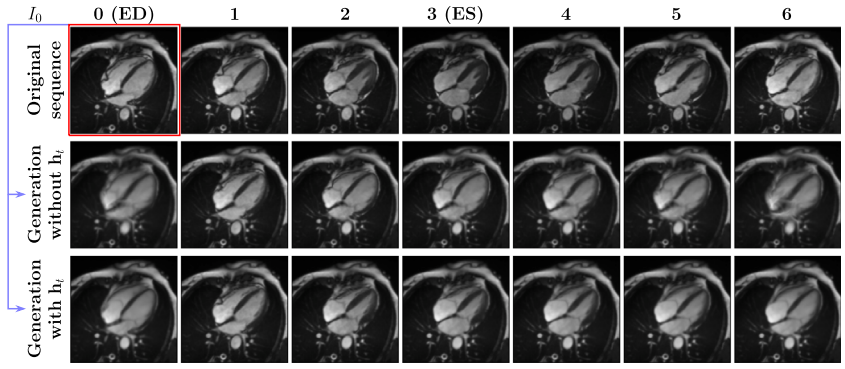

Fig. 9. The effect of hidden state variables of the LSTM on the generation of cardiac MRI sequences. Modelling temporal dependencies via hidden state variables of LSTM ($\mathbf{h}_t$) allows the model to learn the realistic motion of cardiac cycles. In contrast, the model cannot generate meaningful temporal variations without $\mathbf{h}_t$.

of the deformation fields over both space and time allows for efficient motion sampling as well as quantifying the uncertainties without opting for numerical dropout based approaches. The proposed model can compute the motion fields both between the consecutive frames or a reference and other frames, resulting in larger deformations.

In the proposed DragNet framework, we model deformations retroactively using information from the previous time steps through formulating the joint distribution $p(\mathbf{I}_{\leq T-1}, \mathbf{z}_{\leq T-1}, \mathbf{D}_{\leq T-1})$ using a chain rule presented in Eq. (1). We use an LSTM network to model the retrospective dependencies to discern the temporal variations on the latent space and deformations. The hidden state variables of the LSTM update recursively from the current data and the previous hidden states (i.e., $\mathbf{h}_{t-1}$ to keep track of the past) via Eq. (9). The hidden state variables play a key role in retrieving temporal dependencies between the motion fields. These variables convey the information from previous deformations in the sequence to inform the next deformation

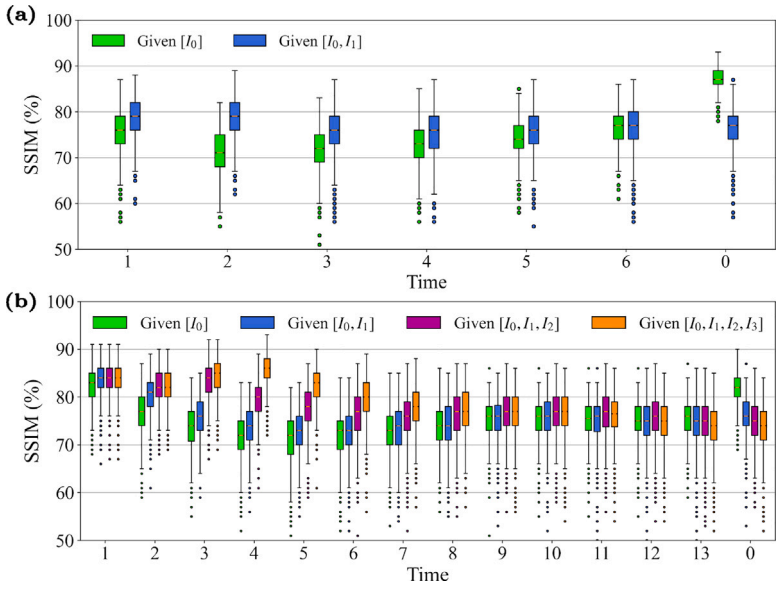

Fig. 10. Assessment of the image quality in 62 000 synthetic sequences in terms of structural similarity index measure (SSIM) between the generated and the original test samples in (a) the seven-frame model (b) fourteen-frame model. The results indicate that more frames lead to initial higher SSIM values which tend to decrease towards the final frames.

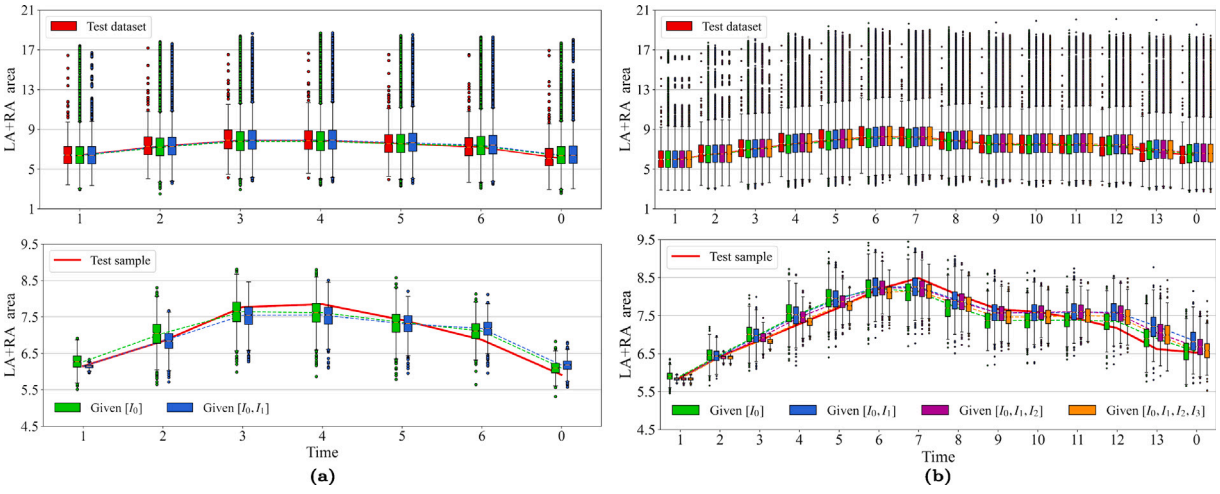

Fig. 11. Assessment of the quality of the motion in the generated sequences via analysing the variations of LA + RA areas extracted by warping the ED masks towards the end of the cycle using the synthetic motion sequences. The results are plotted along with the variations seen in the original test population. Top row: 62K synthetic sequences are compared with the 620 actual test sequences. Bottom row: 100 generated motion sequences from a random test sample are compared with the real cardiac motion. Results in the seven-frame (a) and fourteen-frame model (b). The effect of the model observing more than one frame to initiate the generation process is also shown. The synthetic motion sequences result in patterns similar to those observed in the original test samples, showing diversity.

Table 3
Results of registering the ED phase to other frames quantifying the RMSE, the numbers of negative Jacobian determinant pixels, Dice score between the registered LA+RA and the reference masks, and temporal gradients of the DVFs shown in mean ± std values. **Bold** values indicate the best results among different techniques in each column..

Method	RMSE	$\#\lvert J_{\mathbf{D}}\rvert \leq 0$	Dice (%)	Temp.Grad.
Demons	0.046 ± 0.01	**7755**	76.7 ± 7.3	0.52 ± 0.15
ANTs SyN	0.048 ± 0.01	18 598	80.1 ± 5.9	0.41 ± 0.08
NiftyReg	0.055 ± 0.01	28 914	80.5 ± 7.2	0.67 ± 0.16
VoxelMorph	**0.041 ± 0.01**	27 917	82.2 ± 4.5	0.33 ± 0.06
3D-Elastix	0.049 ± 0.01	31 106	**83.3 ± 4.3**	0.23 ± 0.05
DragNet	0.044 ± 0.01	18 940	82.4 ± 3.8	**0.21 ± 0.04**

field, which is more efficient than random initialisation and coarse-to-fine tuning. Because LSTM matches our modelling formulation and our objective to generate high temporal resolution sequences while capturing long-distance dependencies, we have chosen it as a solution. However, an alternative approach such as temporal convolutional network (TCN) has been used by Krebs et al. (2021) is beneficial for focusing on local dependencies and short-time sequences.

The prior and posterior probability distributions of the latent variables $\mathbf{z}_t$ at time t are conditioned on the hidden states of the recurrent network, $\mathbf{h}_{t-1}$, which include temporal dependencies of previous displacements. This modelling improves the temporal regularisation and precision of the posteriors of $\mathbf{z}_t$ and, therefore, the deformations $\mathbf{D}_t$, resulting in a registration error equivalent to or a bit higher than that of a UNet-based framework like VoxelMorph (Figs. 12 and 13). We further condition the posterior of the motion fields on features from the moving image, which can also help retrieve details. In contrast to Dalca et al. (2018), where only the velocity fields are explicitly modelled via Gaussians at the pixel level, we explicitly define pixel-wise distributions for the displacement fields. Modelling the variances of displacements via $\boldsymbol{\Sigma}_{\mathbf{D}_t}$ variables eliminates the need for computing

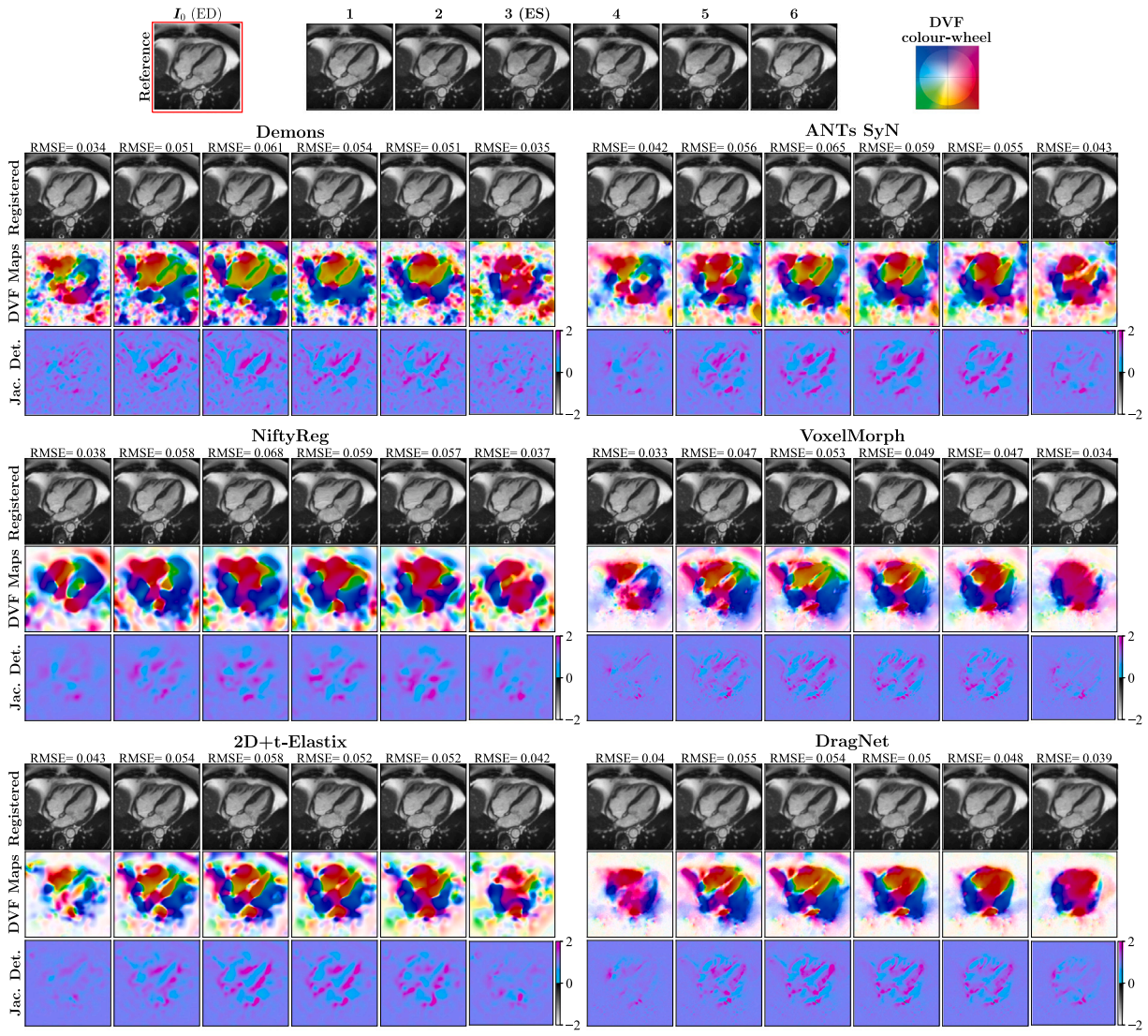

Fig. 12. Registration of the ED phase to other cardiac phases, generating large displacement fields. The results are presented for different registration techniques. The registered images by DragNet are in good agreement with the target images and comparable with the state-of-the-art registration methods. The corresponding colour-coded DVF maps are spatio-temporally smooth in DragNet and have positive Jacobian determinants. The reference CMR images were reproduced with a permission of UK Biobank©.

the registration uncertainty maps via empirical sampling (as proposed in Dalca et al. (2018)) for velocities and propagating those via diffeomorphic layers, which can be computationally costly. It should be noted that, the UNet-based registration models such as Dalca et al. (2018) and VoxelMorph (Balakrishnan et al., 2019), capture image details through skip connections in these structures, resulting in a higher displacement detail level and, consequently, a higher registration accuracy. However, from the generation perspective and diversity of the generated samples (not explored in Dalca et al. (2018)), relying on the latent variables is advantageous in our work even if at the cost of losing some details. Here, we assume similar image intensity distributions for scans, however other similarity metrics, such as cross-correlation or mutual information between the warped and target images, can also be used as the registration error.

The proposed model can also be used to generate a large number of synthetic CMR sequences. We demonstrated this using a single frame as the initial frame. We showed that feeding the network with more than one frame leads to more similarity to the source motion in the first frames but more deviation at the ending frames, especially in long sequences.

Comparing the model performance with five other deformable registration techniques in Tables 1 and 3 showed that the DragNet registration accuracy is comparable to VoxelMorph and generally better than others while improving the spatio-temporal smoothness of the derived displacements. It should be noted that the DragNet is a generative model, while VoxelMorph is a deterministic learning-based approach based on the UNet structure and not a generative model. DragNet is considerably faster than conventional registration techniques such as Demons, ANTs, NiftyReg, and 2D + t-Elastix in computing the entire sequence of motion fields. Training the DragNet takes 72 min on 4K seven-frame CMR sequences using an NVIDIA GeForce RTX 3080 GPU. However, it takes only 0.042 s to compute the entire DVF maps per test sequence. Table 4 compares the runtime for different methods on the GPU and an Intel(R) Core(TM)i9-10850K CPU for a seven-frame sample test sequence.

5. Conclusion

In this paper, we developed a probabilistic model for cardiac cine CMR registration and deriving spatio-temporal deformations featuring a fast runtime. The model is generative and can efficiently generate a large number of synthetic motion sequences. Compared to the

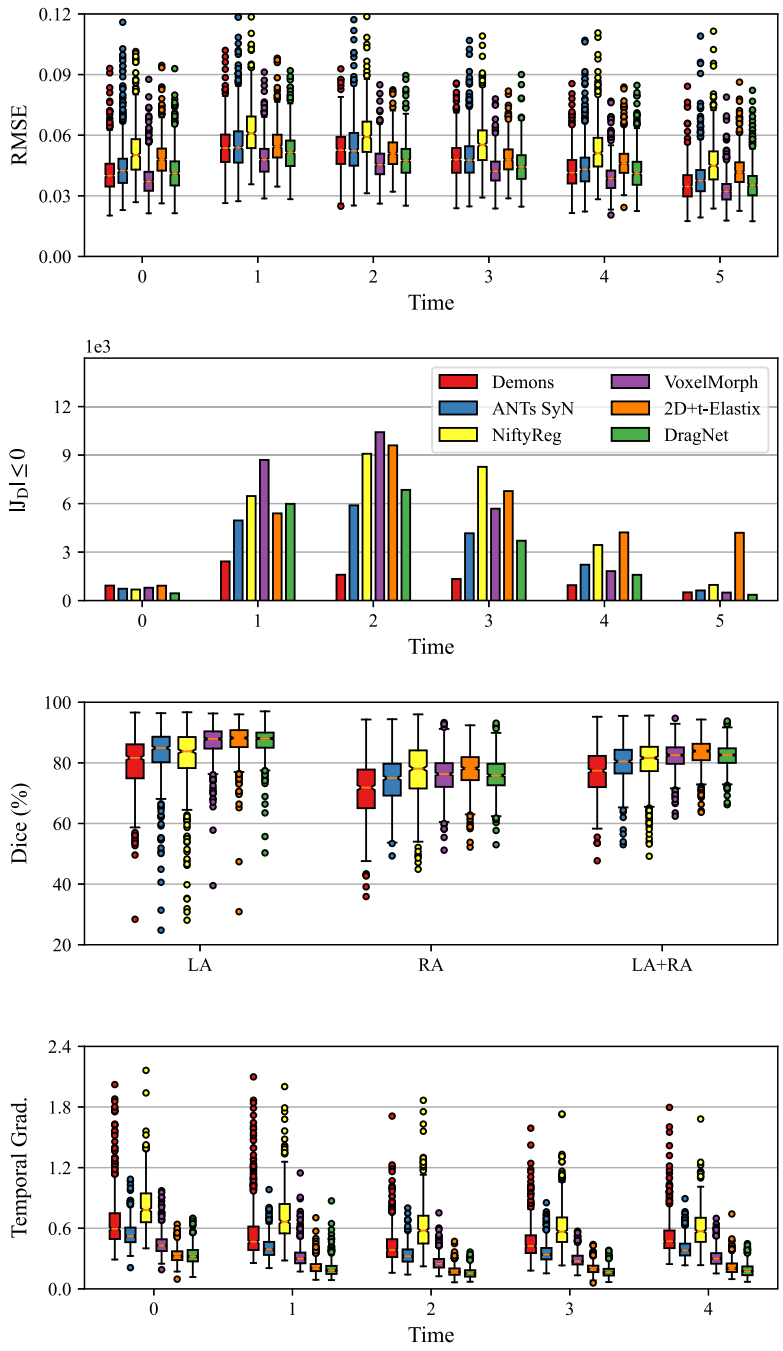

Fig. 13. The results of spatio-temporal registration between frame ED and other frames in the seven-frame CMR sequences showing RMSE, the number of non-positive Jacobian determinant locations, the Dice scores for anatomical structures at ES, and the temporal gradients of the displacement fields. DragNet shows comparable RMSE results to those from VoxelMorph, but smoother DVFs and improves temporal gradients compared with all methods.

Table 4
Runtimes for different techniques to compute DVFs from an unseen seven-frame test sequence.

Method	GPU (s)	CPU (s)
Demons	–	4.586
ANTs SyN	–	15.29
NiftyReg	–	1.55
VoxelMorph	0.034	0.086
2D+t-Elastix	–	5.35
DragNet	0.042	0.21

conventional image registration techniques, the proposed model is significantly faster and capable of generating smoother temporal deformations. DragNet also computes explicit form of spatio-temporal uncertainty estimates, making the results more accountable for the clinical procedures. In terms of speed, it derives the DVFs of a new CMR image sequence in under a second. This feature makes the proposed model more applicable in real-time settings, such as radiotherapy applications where fast compensating for tumour displacement due to breathing becomes crucial. Besides, the proposed framework synthesises high temporal resolution cardiac motion sequences, which can be applicable for recovering missing frames in a cardiac sequence, validation of supervised DIR algorithms, and even enabling in-silico trials which involve modelling a specific moving organ. Our future

work will focus on extending this model to automated abnormal cardiac motion detection considering patient metadata. Besides, we will extend the proposed framework to 3D + t settings suitable for applications such as respiratory motion modelling and 4D-CT data analysis.

CRediT authorship contribution statement

Arezoo Zakeri: Conceptualization, Methodology, Investigation, Software, Formal analysis, Writing – original draft. **Alireza Hokmabadi:** Conceptualization, Methodology, Investigation, Software, Formal analysis, Writing – original draft. **Ning Bi:** Editing. **Isuru Wijesinghe:** Editing. **Michael G. Nix:** Writing – review & editing. **Steffen E. Petersen:** Resources. **Alejandro F. Frangi:** Resources, Writing – review & editing. **Zeike A. Taylor:** Conceptualization, Funding acquisition, Writing – review & editing, Supervision. **Ali Gooya:** Conceptualization, Resources, Funding acquisition, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Steffen E Petersen reports a relationship with Circle Cardiovascular Imaging Inc that includes: consulting or advisory and equity or stocks.

Data availability

The authors do not have permission to share data.

Acknowledgements

The authors would like to acknowledge Cancer Research UK funding (grant C19942/A28832) for the Leeds Radiotherapy Research Centre of Excellence (RadNet). This work was also supported by the Engineering and Physical Sciences Research Council (EPSRC) (grant number EP/S012796/1). ZAT is supported by an RAEng/Leverhulme Trust Research Fellowship (DIADEM-ART LTRF2021-17115). AFF is supported by the RAEng Chair in Emerging Technologies (INSILEX Programme CiET1819/19). BHF PG/14/89/31194, NIHR EP/P001009/1, AI4VBH Center, CAP-AI programme supporting SEP. The UKB CMR dataset have been provided under UK Biobank access application number 11 350. See the Supplementary Material for Ethics and Data availability statements. The authors thank all UKB participants and staff.

Appendix A. Computing $\mathcal{L}_{\mathbf{z}}$

$$
\begin{aligned}
\mathcal{L}_{\mathbf{z}} &= \mathbb{E}_{\prod_{t=0}^{T-1} q(\mathbf{z}_t|\mathbf{I}_t,\mathbf{h}_{t-1})q(\mathbf{D}_t|\mathbf{I}_{t-1},\mathbf{z}_t)} \sum_{t=0}^{T-1} \log \frac{p(\mathbf{z}_t|\mathbf{h}_{t-1})}{q(\mathbf{z}_t|\mathbf{I}_t,\mathbf{h}_{t-1})} \\
&= -\sum_{t=0}^{T-1} \int_{\mathbf{z}_t}\int_{\mathbf{D}_t} q(\mathbf{D}_t|\mathbf{I}_{t-1},\mathbf{z}_t)(q(\mathbf{z}_t|\mathbf{I}_t,\mathbf{h}_{t-1}) \log \frac{q(\mathbf{z}_t|\mathbf{I}_t,\mathbf{h}_{t-1})}{p(\mathbf{z}_t|\mathbf{h}_{t-1})} d\mathbf{z}_t d\mathbf{D}_t \\
&= -\sum_{t=0}^{T-1} \int_{\mathbf{z}_t} (q(\mathbf{z}_t|\mathbf{I}_t,\mathbf{h}_{t-1}) \log \frac{q(\mathbf{z}_t|\mathbf{I}_t,\mathbf{h}_{t-1})}{p(\mathbf{z}_t|\mathbf{h}_{t-1})}) \int_{\mathbf{D}_t} q(\mathbf{D}_t|\mathbf{I}_{t-1},\mathbf{z}_t) d\mathbf{D}_t d\mathbf{z}_t \\
&= -\sum_{t=0}^{T-1} \mathcal{D}_{\mathcal{KL}}\Big(q(\mathbf{z}_t|\mathbf{I}_t,\mathbf{h}_{t-1}) \parallel p(\mathbf{z}_t|\mathbf{h}_{t-1})\Big) \\
&= -0.5 \sum_{t=0}^{T-1} (\log \boldsymbol{\sigma}^2_{\mathbf{z}_t} - \log \boldsymbol{\sigma}^2_{\mathbf{z}_t,pi} + 1 - \frac{\boldsymbol{\sigma}^2_{\mathbf{z}_t} + (\boldsymbol{\mu}_{\mathbf{z}_t} - \boldsymbol{\mu}_{\mathbf{z}_t,pi})^2}{\boldsymbol{\sigma}^2_{\mathbf{z}_t,pi}}) \qquad \text{(A.1)}
\end{aligned}
$$

Appendix B. Computing $\mathcal{L}_{\mathbf{D}}$

$$
\begin{aligned}
\mathcal{L}_{\mathbf{D}} &= \mathbb{E}_{\prod_{t=0}^{T-1} q(\mathbf{z}_t|\mathbf{I}_t,\mathbf{h}_{t-1})q(\mathbf{D}_t|\mathbf{I}_{t-1},\mathbf{z}_t)} \sum_{t=0}^{T-1} \log \frac{p(\mathbf{D}_t|\mathbf{h}_{t-1})}{q(\mathbf{D}_t|\mathbf{I}_{t-1},\mathbf{z}_t)} \\
&= -\sum_{t=0}^{T-1} \int_{\mathbf{z}_t}\int_{\mathbf{D}_t} q(\mathbf{z}_t|\mathbf{I}_t,\mathbf{h}_{t-1})q(\mathbf{D}_t|\mathbf{I}_{t-1},\mathbf{z}_t) \log \frac{q(\mathbf{D}_t|\mathbf{I}_{t-1},\mathbf{z}_t)}{p(\mathbf{D}_t|\mathbf{h}_{t-1})} d\mathbf{z}_t d\mathbf{D}_t \\
&= -\sum_{t=0}^{T-1} \int_{\mathbf{z}_t} q(\mathbf{z}_t|\mathbf{I}_t,\mathbf{h}_{t-1}) \Big(\int_{\mathbf{D}_t} q(\mathbf{D}_t|\mathbf{I}_{t-1},\mathbf{z}_t) \log \frac{q(\mathbf{D}_t|\mathbf{I}_{t-1},\mathbf{z}_t)}{p(\mathbf{D}_t|\mathbf{h}_{t-1})} d\mathbf{D}_t \Big) d\mathbf{z}_t \\
&\simeq -\sum_{t=0}^{T-1} \frac{1}{L} \sum_{l=1}^{L} \mathcal{D}_{\mathcal{KL}}\Big(q(\mathbf{D}_t|\mathbf{I}_{t-1},\mathbf{z}_t^{(l)}) \parallel p(\mathbf{D}_t|\mathbf{h}_{t-1})\Big) \qquad \text{(B.1)}
\end{aligned}
$$

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2022.102678.

References

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. Neuroimage 38 (1), 95–113.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12 (1), 26–41.

Avants, B.B., Tustison, N.L., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54 (3), 2033–2044.

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging 38 (8), 1788–1800.

Chen, X., Diaz-Pinto, A., Ravikumar, N., Frangi, A.F., 2021. Deep learning in medical image registration. Prog. Biomed. Eng. 3 (1), 012003.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A.C., Bengio, Y., 2015. A recurrent latent variable model for sequential data. Adv. Neural Inf. Process. Syst. 28.

Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R., 2018. Unsupervised learning for fast probabilistic diffeomorphic registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 729–738.

De Craene, M., Piella, G., Camara, O., Duchateau, N., Silva, E., Doltra, A., D'hooge, J., Brugada, J., Sitges, M., Frangi, A.F., 2012. Temporal diffeomorphic free-form deformation: Application to motion and strain estimation from 3D echocardiography. Med. Image Anal. 16 (2), 427–450.

De Vos, B.D., Berendsen, F.F., Viergever, M.A., Sokooti, H., Staring, M., Išgum, I., 2019. A deep learning framework for unsupervised affine and deformable image registration. Med. Image Anal. 52, 128–143.

De Vos, B.D., Berendsen, F.F., Viergever, M.A., Staring, M., Išgum, I., 2017. End-to-end unsupervised deformable image registration with a convolutional neural network. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 204–212.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T., 2015. Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2758–2766.

Eppenhof, K.A., Lafarge, M.W., Moeskops, P., Veta, M., Pluim, J.P., 2018. Deformable image registration using convolutional neural networks. In: Medical Imaging 2018: Image Processing, Vol. 10574. International Society for Optics and Photonics, p. 105740S.

Eppenhof, K.A., Pluim, J.P., 2018. Pulmonary CT registration through supervised learning with convolutional neural networks. IEEE Trans. Med. Imaging 38 (5), 1097–1105.

Fan, J., Cao, X., Yap, P.-T., Shen, D., 2019. BIRNet: Brain image registration using dual-supervised fully convolutional networks. Med. Image Anal. 54, 193–206.

Gal, Y., Ghahramani, Z., 2015. Bayesian convolutional neural networks with Bernoulli approximate variational inference. arXiv preprint arXiv:1506.02158.

Giger, A., Sandkühler, R., Jud, C., Bauman, G., Bieri, O., Salomir, R., Cattin, P.C., 2018. Respiratory motion modelling using cGANs. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 81–88.

Hershey, J.R., Olsen, P.A., 2007. Approximating the Kullback Leibler divergence between Gaussian mixture models. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, Vol. 4. IEEE, pp. IV–317.

Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks. Adv. Neural Inf. Process. Syst. 28.

Kendall, A., Badrinarayanan, V., Cipolla, R., 2015. Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv preprint arXiv:1511.02680.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. 30.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

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

Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D., Ronneberger, O., 2018. A probabilistic U-net for segmentation of ambiguous images. Adv. Neural Inf. Process. Syst. 31.

Krebs, J., Delingette, H., Ayache, N., Mansi, T., 2021. Learning a generative motion model from image sequences based on a latent motion matrix. IEEE Trans. Med. Imaging 40 (5), 1405–1416.

Krebs, J., Delingette, H., Mailhé, B., Ayache, N., Mansi, T., 2019a. Learning a probabilistic model for diffeomorphic registration. IEEE Trans. Med. Imaging 38 (9), 2165–2176.

Krebs, J., Mansi, T., Ayache, N., Delingette, H., 2019b. Probabilistic motion modeling from medical image sequences: application to cardiac cine-MRI. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 176–185.

Krebs, J., Mansi, T., Mailhé, B., Ayache, N., Delingette, H., 2018. Unsupervised probabilistic deformation modeling for robust diffeomorphic registration. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 101–109.

Kuang, D., Schmah, T., 2019. FAIM–a convnet method for unsupervised 3D medical image registration. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 646–654.

Marstal, K., Berendsen, F., Staring, M., Klein, S., 2016. SimpleElastix: A user-friendly, multi-lingual library for medical image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 134–142.

Metz, C.T., Klein, S., Schaap, M., van Walsum, T., Niessen, W.J., 2011. Nonrigid registration of dynamic medical imaging data using nD+ t B-splines and a groupwise optimization approach. Med. Image Anal. 15 (2), 238–249.

Mezheritsky, T., Romaguera, L.V., Le, W., Kadoury, S., 2022. Population-based 3D respiratory motion modelling from convolutional autoencoders for 2D ultrasound-guided radiotherapy. Med. Image Anal. 75, 102260.

Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S., 2010. Fast free-form deformation using graphics processing units. Comput. Methods Programs Biomed. 98 (3), 278–284.

Petersen, S.E., Matthews, P.M., Francis, J.M., Robson, M.D., Zemrak, F., Boubertakh, R., Young, A.A., Hudson, S., Weale, P., Garratt, S., et al., 2015. UK biobank's cardiovascular magnetic resonance protocol. J. Cardiovasc. Magn. Reson. 18 (1), 1–7.

Psaros, A.F., Meng, X., Zou, Z., Guo, L., Karniadakis, G.E., 2022. Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. arXiv preprint arXiv:2201.07766.

Rohé, M.-M., Sermesant, M., Pennec, X., 2018. Low-dimensional representation of cardiac motion using barycentric subspaces: a new group-wise paradigm for estimation, analysis, and reconstruction. Med. Image Anal. 45, 1–12.

Salehi, S.S.M., Khan, S., Erdogmus, D., Gholipour, A., 2018. Real-time deep registration with geodesic loss. arXiv preprint arXiv:1803.05982.

Sentker, T., Madesta, F., Werner, R., 2018. GDL-FIRE 4D: Deep learning-based fast 4D CT image registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 765–773.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-c., 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Adv. Neural Inf. Process. Syst. 28.

Sokooti, H., de Vos, B., Berendsen, F., Ghafoorian, M., Yousefi, S., Lelieveldt, B.P., Isgum, I., Staring, M., 2019. 3D convolutional neural networks image registration based on efficient supervised learning from artificial deformations. arXiv preprint arXiv:1908.10235.

Teng, X., Chen, Y., Zhang, Y., Ren, L., 2021. Respiratory deformation registration in 4D-CT/cone beam CT using deep learning. Quant. Imaging Med. Surg. 11 (2), 737.

Uzunova, H., Wilms, M., Handels, H., Ehrhardt, J., 2017. Training CNNs for image registration from few samples with model-based data augmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 223–231.

Vercauteren, T., Pennec, X., Perchant, A., Ayache, N., 2008. Symmetric log-domain diffeomorphic registration: A demons-based approach. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 754–761.

Vercauteren, T., Pennec, X., Perchant, A., Ayache, N., 2009. Diffeomorphic demons: Efficient non-parametric image registration. NeuroImage 45 (1), S61–S72.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Wilson, A.G., Izmailov, P., 2020. Bayesian deep learning and a probabilistic perspective of generalization. Adv. Neural Inf. Process. Syst. 33, 4697–4708.

Yu, J.J., Harley, A.W., Derpanis, K.G., 2016. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In: European Conference on Computer Vision. Springer, pp. 3–10.

Zhang, J., 2018. Inverse-consistent deep networks for unsupervised deformable image registration. arXiv preprint arXiv:1809.03443.