DeepTag: An Unsupervised Deep Learning Method for Motion Tracking on Cardiac Tagging Magnetic Resonance Images

Meng Ye[1], Mikael Kanski[2], Dong Yang[3], Qi Chang[1],
Zhennan Yan[4], Qiaoying Huang[1], Leon Axel[2], Dimitris Metaxas[1]

[1]Rutgers University, [2]New York University School of Medicine, [3]NVIDIA,
[4]SenseBrain and Shanghai AI Laboratory and Centre for Perceptual and Interactive Intellgience
{my389, qc58, qh55, dnm}@cs.rutgers.edu

Abstract

Cardiac tagging magnetic resonance imaging (t-MRI) is the gold standard for regional myocardium deformation and cardiac strain estimation. However, this technique has not been widely used in clinical diagnosis, as a result of the difficulty of motion tracking encountered with t-MRI images. In this paper, we propose a novel deep learning-based fully unsupervised method for in vivo motion tracking on t-MRI images. We first estimate the motion field (INF) between any two consecutive t-MRI frames by a bi-directional generative diffeomorphic registration neural network. Using this result, we then estimate the Lagrangian motion field between the reference frame and any other frame through a differentiable composition layer. By utilizing temporal information to perform reasonable estimations on spatio-temporal motion fields, this novel method provides a useful solution for motion tracking and image registration in dynamic medical imaging. Our method has been validated on a representative clinical t-MRI dataset; the experimental results show that our method is superior to conventional motion tracking methods in terms of landmark tracking accuracy and inference efficiency. Project page is at: https://github.com/DeepTag/cardiac_tagging_motion_estimation.

1. Introduction

Cardiac magnetic resonance imaging (MRI) provides a non-invasive way to evaluate the morphology and function of the heart from the imaging data. Specifically, dynamic cine imaging, which generates a 2D image sequence to cover a full cardiac cycle, can provide direct information of heart motion. Due to the long imaging time and breath-holding requirements, the clinical cardiac MRI imaging protocols are still 2D sequences. To recover the 3D motion field of the whole heart wall, typically we need to scan several slices in long axis (2-, 3-, 4-chamber) views and short-axis (SAX) views, as shown in Fig. 1. There are two kinds of dynamic imaging: conventional (untagged) cine MR imaging and tagging imaging (t-MRI) [1]. For untagged cine images (most recent work has focused on these images), feature tracking can be used to estimate myocardial motion [22, 35, 40, 57, 55, 54]. However, as shown in Fig. 1 (b), due to the relatively uniform signal in the myocardial wall and the lack of reliable identifiable landmarks, the estimated motion cannot be used as a reliable indicator for clinical diagnosis. In contrast, t-MRI provides the gold standard imaging method for regional myocardial motion quantification and strain estimation. The t-MRI data is produced by a specially designed magnetic preparation module

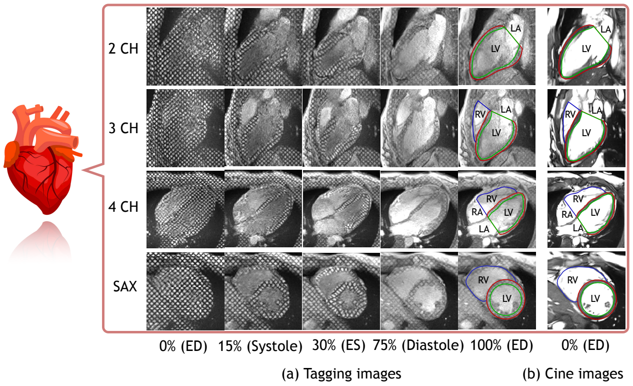

Figure 1. Standard scan views (2-, 3-, 4-chamber views and short-axis views) of cardiac MRI. (a) Tagging images. Number under the figure means percentage of one cardiac cycle. (b) End-diastole (ED) phase of cine images. Red and green contours depict the epi- and endo-cardial borders of left ventricle (LV) myocardium (MYO) wall. Blue contour depicts the right ventricle (RV). LA: left atrium. RA: right atrium.

called spatial modulation of magnetization (SPAMM) [5]. It introduces the intrinsic tissue markers which are stripe-like darker tag patterns embedded in relatively brighter myocardium, as shown in Fig. 1 (a). By tracking the deformation of tags, we can retrieve a 2D displacement field in the imaging plane and recover magnetization, which non-invasively creates fiducial "tags" within the heart wall.

Although it has been widely accepted as the gold standard imaging modality for regional myocardium motion quantification, t-MRI has largely remained only a research tool and has not been widely used in clinical practice. The principal challenge (detailed analysis in Supplementary Material) is the associated time-consuming post-processing, which could be principally attributed to the following: (1) Image appearance changes greatly over a cardiac cycle and tag signal fades on the later frames, as shown in Fig. 1 (a). (2) Motion artifacts can degrade images. (3) Other artifacts and noise can reduce image quality. To tackle these problems, in this work, we propose a novel deep learning-based unsupervised method to estimate tag deformations on t-MRI images. The method has no annotation requirement during training, so with more training data are collected, our method can learn to predict more accurate cardiac deformation motion fields with minimal increased effort. In our method, we first track the motion field in between two consecutive frames, using a bi-directional generative diffeomorphic registration network. Based on these initial motion field estimations, we then track the Lagrangian motion field between the reference frame and any other frame by a composition layer. The composition layer is differentiable, so it can update the learning parameters of the registration network with a global Lagrangian motion constraint, thus achieving a reasonable computation of motion fields.

Our contributions could be summarized briefly as follows: (1) We propose a novel unsupervised method for t-MRI motion tracking, which can achieve a high accuracy of performance in a fast inference speed. (2) We propose a bi-directional diffeomorphic image registration network which could guarantee topology preservation and invertibility of the transformation, in which the likelihood of the warped image is modeled as a Boltzmann distribution, and a normalized cross correlation metric is incorporated in it, for its robust performance on image intensity time-variant registration problems. (3) We propose a scheme to decompose the Lagrangian motion between the reference and any other frame into sums of consecutive frame motions and then improve the estimation of these motions by composing them back into the Lagrangian motion and posing a global motion constraint.

2. Background

Regional myocardium motion quantification mainly focuses on the left ventricle (LV) myocardium (MYO) wall. It takes one t-MRI image sequence (usually a 2D video) as input and outputs a 2D motion field over time. The motion field is a 2D dense field depicting the non-rigid deformation of the LV MYO wall. The image sequence covers a full cardiac cycle. It starts from the end diastole (ED) phase, at which the ventricle begins to contract, then to the maximum contraction at end systole (ES) phase and back to relaxation to ED phase, as shown in Fig. 1. Typically, we set a reference frame as the ED phase, and track the motion on any other later frame relative to the reference one. For t-MRI motion tracking, previous work was mainly based on phase, optical flow, and conventional non-rigid image registration.

2.1. Phase-based Method

Harmonic phase (HARP) based method is the most representative one for t-MRI image motion tracking [37, 38, 28, 27, 17]. Periodic tags in the image domain correspond to spectral peaks in the Fourier domain of the image. Isolating the first harmonic peak region by a bandpass filter and performing an inverse Fourier transform of the selected region yields a complex harmonic image. The phase map of the complex image is the HARP image, which could be used for motion tracking since the harmonic phase of a material point is a time-invariant physics property, for simple translation. Thus, by tracking the harmonic phase vector of each pixel through time, one can track the position and, by extension, the displacement of each pixel along time. However, due to cardiac motion, local variations of tag spacing and orientation at different frames may lead to erroneous phase estimation when using HARP, such as bifurcations in the reconstructed phase map, which also happens at boundaries and in large deformation regions of the myocardium [28]. Extending HARP, Gabor filters are used to refine phase map estimation by changing the filter parameters according to the local tag spacing and orientation, to automatically match different tag patterns in the image domain [13, 50, 39].

2.2. Optical Flow Approach

While HARP exploits specificity of quasiperiodic t-MRI, the optical flow (OF) based method is generic and can be applied to track objects in video sequences [18, 8, 7, 32, 52]. OF can estimate a dense motion field based on the basic assumption of image brightness constancy of local time-varying image regions with motion, at least for a very short time interval. The under-determined OF constraint equation is solved by variational principles in which some other regularization constraints are added in, including the image gradient, the phase or block matching. Although efforts have been made to seek more accurate regularization terms, OF approaches lack accuracy, especially for t-MRI motion tracking, due to the tag fading and large deformation problems [11, 49]. More recently, convolutional neural networks (CNN) are trained to predict OF [16, 19, 20, 24, 26, 41, 31, 47, 53, 51, 48]. However, most of these works were supervised methods, with the need of a ground truth OF for training, which is nearly impossible to obtain for medical images.

2.3. Image Registration-based Method

Conventional non-rigid image registration methods have been used to estimate the deformation of the myocardium for a long time [46, 43, 30, 12, 34, 25]. Non-rigid registration schemes are formulated as an optimization procedure that maximizes a similarity criterion between the fixed image and the transformed moving image, to find the optimal transformation. Transformation models could be parametric models, including B-spline free-form deformation [46, 34, 12], and non-parametric models, including the variational method. Similarity criteria are generally chosen, such as mutual information and generalized information measures [43]. All of these models are iteratively optimized, which is time consuming.

Recently, deep learning-based methods have been applied to medical image registration and motion tracking. They are fast and have achieved at least comparable accuracy with conventional registration methods. Among those approaches, supervised methods [42] require ground truth deformation fields, which are usually synthetic. Registration accuracy thus will be limited by the quality of synthetic ground truth. Unsupervised methods [9, 10, 23, 22, 56, 15, 6, 14, 36, 44, 45, 33] learn the deformation field by a loss function of the similarity between the fixed image and warped moving image. Unsupervised methods have been extended to cover deformable and diffeomorphic models. Deformable models [6, 9, 10] aim to learn the single directional deformation field from the fixed image to the moving image. Diffeomorphic models [14, 22, 33, 45] learn the stationary velocity field (SVF) and integrate the SVF by a scaling and squaring layer, to get the diffeomorphic deformation field [14]. A deformation field with diffeomorphism is differentiable and invertible, which ensures one-to-one mapping and preserves topology. Inspired by these works, we propose to use a bi-directional diffeomorphic registration network to track motions on t-MRI images.

3. Method

We propose an unsupervised learning method based on deep learning to track dense motion fields of objects that change over time. Although our method can be easily extended to other motion tracking tasks, without loss of generality, the design focus of the proposed method is t-MRI motion tracking.

3.1. Motion Decomposition and Recomposition

As shown in Fig. 2, for a material point m which moves from position X_0 at time t_0, we have its trajectory X_t.

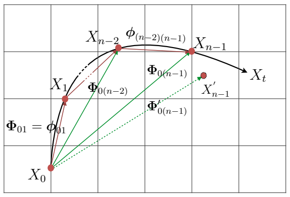

Figure 2. Interframe (INF) motion ϕ and Lagrangian motion $\mathbf{\Phi}$.

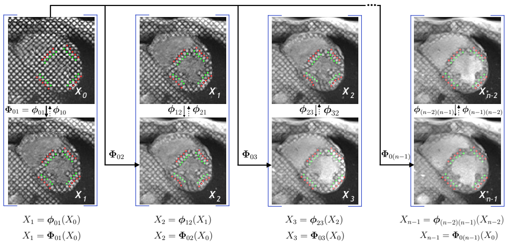
X_0
X_1
X_2
X_{n-2}
$\mathbf{\Phi}_{01} = \phi_{01} \downarrow \uparrow \phi_{10}$
$\phi_{12} \downarrow \uparrow \phi_{21}$
$\phi_{23} \downarrow \uparrow \phi_{32}$
$\phi_{(n-2)(n-1)} \downarrow \uparrow \phi_{(n-1)(n-2)}$
$\mathbf{\Phi}_{02}$
$\mathbf{\Phi}_{03}$
$\mathbf{\Phi}_{0(n-1)}$
X_1
X_2
X_3
X_{n-1}
$X_1 = \phi_{01}(X_0)$
$X_1 = \mathbf{\Phi}_{01}(X_0)$
$X_2 = \phi_{12}(X_1)$
$X_2 = \mathbf{\Phi}_{02}(X_0)$
$X_3 = \phi_{23}(X_2)$
$X_3 = \mathbf{\Phi}_{03}(X_0)$
$X_{n-1} = \phi_{(n-2)(n-1)}(X_{n-2})$
$X_{n-1} = \mathbf{\Phi}_{0(n-1)}(X_0)$

Figure 3. An overview of our scheme for regional myocardium motion tracking on t-MRI image sequences. ϕ: Interframe (INF) motion field between consecutive image pairs. $\mathbf{\Phi}$: Lagrangian motion field between the first frame and any other later frame.

In a N frames sequence, we only record the finite positions $X_n (n = 0, 1, ..., N-1)$ of m. In a time interval $\Delta t = t_{n-1} - t_{n-2}$, the displacement can be shown pictorially as a vector $\phi_{(n-2)(n-1)}$, which in our work we call the interframe (INF) motion. A set of INF motions $\left\{ \phi_{t(t+1)} (t = 0, 1, ..., n-2) \right\}$ will recompose the motion vector $\mathbf{\Phi}_{0(n-1)}$, which we call the Lagrangian motion. While INF motion $\phi_{t(t+1)}$ in between two consecutive frames is small if the time interval Δt is small, net Lagrangian motion $\mathbf{\Phi}_{0(n-1)}$, however, could be very large in some frames of the sequence. For motion tracking, as we set the first frame as the reference frame, our task is to derive the Lagrangian motion $\mathbf{\Phi}_{0(n-1)}$ on any other later frame $t = n-1$. It is possible to directly track it based on the associated frame pairs, but for large motion, the tracking result $\mathbf{\Phi}'_{0(n-1)}$ could drift a lot. In a cardiac cycle, for a given frame $t = n-1$, since the amplitude $\| \phi_{(n-2)(n-1)} \| \leq \| \mathbf{\Phi}_{0(n-1)} \|$, decomposing $\mathbf{\Phi}_{0(n-1)}$ into $\left\{ \phi_{t(t+1)} (t = 0, 1, ..., n-2) \right\}$, tracking $\left\{ \phi_{t(t+1)} \right\}$ at first, then composing them back to $\mathbf{\Phi}_{0(n-1)}$ will make sense. In this work, we follow this idea to obtain accurate motion tracking results on t-MRI images.

3.2. Motion Tracking on A Time Sequence

Fig. 3 shows our scheme for myocardium motion tracking through time on a t-MRI image sequence. We first estimate the INF motion field ϕ between two consecutive frames by a bi-directional diffeomorphic registration network, as shown in Fig. 4. Once all the INF motion

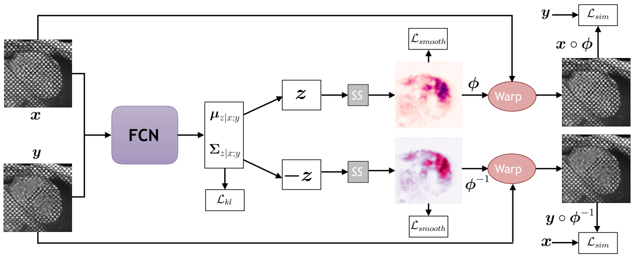

Figure 4. An overview of our proposed bi-directional forward-backward generative diffeomorphic registration network.

fields are obtained in the full time sequence, we compose them as the Lagrangian motion field $\mathbf{\Phi}$, which is shown in Fig. 5. Motion tracking is achieved by predicting the position X_{n-1} on an arbitrary frame moved from the position X_0 on the first frame with the estimated Lagrangian motion field: $X_{n-1} = \mathbf{\Phi}_{0(n-1)}(X_0)$. In our method, motion composition is implemented by a differentiable composition layer C, as depicted in Fig. 6. When training the registration network, such a differentiable layer can backpropagate the similarity loss between the warped reference image by Lagrangian motion field $\mathbf{\Phi}$ and any other later frame image as a global constraint and then update the parameters of the registration net, which in turn guarantees a reasonable INF motion field $\boldsymbol{\phi}$ estimation.

3.3. Bi-Directional Forward-Backward Generative Diffeomorphic Registration Network

As shown in Fig. 4, we use a bi-directional forward-backward diffeomorphic registration network to estimate the INF motion field $\boldsymbol{\phi}$. Our network is modeled as a generative stochastic variational autoencoder (VAE) [21]. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be a 2D image pair, and let $\boldsymbol{z}$ be a latent variable that parameterizes the INF motion field $\boldsymbol{\phi} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$. Following the methodology of a VAE, we assume that the prior $p(\boldsymbol{z})$ is a multivariate Gaussian distribution with zero mean and covariance $\mathbf{\Sigma}_z$:

$$p(\boldsymbol{z}) \sim \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \mathbf{\Sigma}_z). \tag{1}$$

The latent variable $\boldsymbol{z}$ could be applied to a wide range of representations for image registration. In our work, in order to obtain a diffeomorphism, we let $\boldsymbol{z}$ be a SVF which is generated as the path of diffeomorphic deformation field $\boldsymbol{\phi}^{(t)}$ parametrized by $t \in [0, 1]$ as follows:

$$\frac{d\boldsymbol{\phi}^{(t)}}{dt} = \boldsymbol{v}(\boldsymbol{\phi}^{(t)}) = \boldsymbol{v} \circ \boldsymbol{\phi}^{(t)}, \tag{2}$$

where $\circ$ is a composition operator, $\boldsymbol{v}$ is the velocity field ($\boldsymbol{v} = \boldsymbol{z}$) and $\boldsymbol{\phi}^{(0)} = Id$ is an identity transformation. We follow [2, 3, 14, 33] to integrate the SVF $\boldsymbol{v}$ over time $t = [0, 1]$ by a scaling and squaring layer (SS) to obtain the final motion field $\boldsymbol{\phi}^{(1)}$ at time $t = 1$. Specifically, starting from $\boldsymbol{\phi}^{(1/2^T)} = \boldsymbol{p} + \boldsymbol{v}(\boldsymbol{p})/2^T$ where $\boldsymbol{p}$ is a spatial location, by using the recurrence $\boldsymbol{\phi}^{(1/2^t)} = \boldsymbol{\phi}^{(1/2^{t+1})} \circ \boldsymbol{\phi}^{(1/2^{t+1})}$ we can compute $\boldsymbol{\phi}^{(1)} = \boldsymbol{\phi}^{(1/2)} \circ \boldsymbol{\phi}^{(1/2)}$. In our experiments, $T = 7$, which is chosen so that $\boldsymbol{v}(\boldsymbol{p})/2^T$ is small enough. With the latent variable $\boldsymbol{z}$, we can compute the motion field $\boldsymbol{\phi}$ by the SS layer. We then use a spatial transform layer to warp image $\boldsymbol{x}$ by $\boldsymbol{\phi}$ and we obtain a noisy observation of the warped image, $\boldsymbol{x} \circ \boldsymbol{\phi}$, which could be a Gaussian distribution:

$$p(\boldsymbol{y}|\boldsymbol{z}; \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{x} \circ \boldsymbol{\phi}, \sigma^2 \mathbb{I}), \tag{3}$$

where $\boldsymbol{y}$ denotes the observation of warped image $\boldsymbol{x}$, σ^2 describes the variance of additive image noise. We call the process of warping image $\boldsymbol{x}$ towards $\boldsymbol{y}$ as the forward registration.

Our goal is to estimate the posterior probabilistic distribution $p(\boldsymbol{z}|\boldsymbol{y}; \boldsymbol{x})$ for registration so that we obtain the most likely motion field $\boldsymbol{\phi}$ for a new image pair $(\boldsymbol{x}, \boldsymbol{y})$ via maximum a posteriori estimation. However, directly computing this posterior is intractable. Alternatively, we can use a variational method, and introduce an approximate multivariate normal posterior probabilistic distribution $q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y}; \boldsymbol{x})$ parametrized by a fully convolutional neural network (FCN) module $\boldsymbol{\psi}$ as:

$$q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y}; \boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{z|x,y}, \mathbf{\Sigma}_{z|x,y}), \tag{4}$$

where we let the FCN learn the mean $\boldsymbol{\mu}_{z|x,y}$ and diagonal covariance $\mathbf{\Sigma}_{z|x,y}$ of the posterior probabilistic distribution $q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y}; \boldsymbol{x})$. When training the network, we implement a layer that samples a new latent variable $\boldsymbol{z}_k$ using the reparameterization trick: $\boldsymbol{z}_k = \boldsymbol{\mu}_{z|x,y} + \epsilon \mathbf{\Sigma}_{z|x,y}$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$.

To learn parameters $\boldsymbol{\psi}$, we minimize the KL divergence between $q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y}; \boldsymbol{x})$ and $p(\boldsymbol{z}|\boldsymbol{y}; \boldsymbol{x})$, which leads to maximizing the evidence lower bound (ELBO) [21] of the log marginalized likelihood $log\ p(\boldsymbol{y}|\boldsymbol{x})$, as follows (detailed derivation in Supplementary Material):

$$\begin{aligned} &\underset{\boldsymbol{\psi}}{min}\ \mathcal{KL}[q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y}; \boldsymbol{x}) || p(\boldsymbol{z}|\boldsymbol{y}; \boldsymbol{x})] \\ &= \underset{\boldsymbol{\psi}}{min}\ \mathcal{KL}[q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y}; \boldsymbol{x}) || p(\boldsymbol{z})] - \mathbb{E}_q[log\, p(\boldsymbol{y}|\boldsymbol{z}; \boldsymbol{x})] \\ &+ log\ p(\boldsymbol{y}|\boldsymbol{x}). \end{aligned} \tag{5}$$

In Eq. (5), the second term $-\mathbb{E}_q[log\, p(\boldsymbol{y}|\boldsymbol{z}; \boldsymbol{x})]$ is called the reconstruction loss term in a VAE model. While we can model the distribution of $p(\boldsymbol{y}|\boldsymbol{z}; \boldsymbol{x})$ as a Gaussian as in Eq. (3), which is equivalent to using a sum-of-squared difference (SSD) metric to measure the similarity between the warped image $\boldsymbol{x}$ and the observed $\boldsymbol{y}$, in this work, we instead use a normalized local cross-correlation (NCC) metric, due to its robustness properties and superior results, especially for intensity time-variant image registration problems [4, 29]. NCC of an image pair I and J is defined as:

$$NCC(I, J) = \sum_{\boldsymbol{p} \in \Omega} \frac{\sum_{\boldsymbol{p}_i} (I(\boldsymbol{p}_i) - \bar{I}(\boldsymbol{p}))(J(\boldsymbol{p}_i) - \bar{J}(\boldsymbol{p}))}{\sqrt{\sum_{\boldsymbol{p}_i} (I(\boldsymbol{p}_i) - \bar{I}(\boldsymbol{p}))^2 \sum_{\boldsymbol{p}_i} (J(\boldsymbol{p}_i) - \bar{J}(\boldsymbol{p}))^2}}, \quad (6)$$

where $\bar{I}(\boldsymbol{p})$ and $\bar{J}(\boldsymbol{p})$ are the local mean of I and J at position $\boldsymbol{p}$ respectively calculated in a w^2 window Ω centered at $\boldsymbol{p}$. In our experiments, we set $w = 9$. A higher NCC indicates a better alignment, so the similarity loss between I and J could be: $\mathcal{L}_{sim}(I, J) = -NCC(I, J)$. Thus, we adopt the following Boltzmann distribution to model $p(\boldsymbol{y}|\boldsymbol{z}; \boldsymbol{x})$ as:

$$p(\boldsymbol{y}|\boldsymbol{z}; \boldsymbol{x}) \sim exp(-\gamma NCC(\boldsymbol{y}, \boldsymbol{x} \circ \boldsymbol{\phi})), \quad (7)$$

where γ is a negative scalar hyperparameter. Finally, we formulate the loss function as:

$$\begin{aligned} \mathcal{L}_{kl} &= \mathcal{KL}[q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y}; \boldsymbol{x})||p(\boldsymbol{z})] - \mathbb{E}_q[log\, p(\boldsymbol{y}|\boldsymbol{z}; \boldsymbol{x})] + const \\ &= \frac{1}{2}\left[tr(\lambda D \boldsymbol{\Sigma}_{z|x,y} - log \boldsymbol{\Sigma}_{z|x,y}) + \boldsymbol{\mu}_{z|x,y}^T \boldsymbol{\Lambda}_z \boldsymbol{\mu}_{z|x,y}\right] \\ &+ \frac{\gamma}{K} \sum_k NCC(\boldsymbol{y}, \boldsymbol{x} \circ \boldsymbol{\phi}_k) + const, \end{aligned} \quad (8)$$

where $\boldsymbol{D}$ is the graph degree matrix defined on the 2D image pixel grid and K is the number of samples used to approximate the expectation, with $K = 1$ in our experiments. We let $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$ be the Laplacian of a neighborhood graph defined on the pixel grid, where $\boldsymbol{A}$ is a pixel neighborhood adjacency matrix. To encourage the spatial smoothness of SVF $\boldsymbol{z}$, we set $\boldsymbol{\Lambda}_z = \boldsymbol{\Sigma}_z^{-1} = \lambda \boldsymbol{L}$ [14], where λ is a parameter controlling the scale of the SVF $\boldsymbol{z}$.

With the SVF representation, we can also compute an inverse motion field $\boldsymbol{\phi}^{-1}$ by inputting $-\boldsymbol{z}$ into the SS layer: $\boldsymbol{\phi}^{-1} = SS(-\boldsymbol{z})$. Thus we can warp image $\boldsymbol{y}$ towards image $\boldsymbol{x}$ (the backward registration) and get the observation distribution of warped image $\boldsymbol{y}$: $p(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{y})$. We minimize the KL divergence between $q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{y})$ and $p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{y})$ which leads to maximizing the ELBO of the log marginalized likelihood $log\, p(\boldsymbol{x}|\boldsymbol{y})$ (see supplementary material for detailed derivation). In this way, we can add the backward KL loss term into the forward KL loss term and get:

$$\begin{aligned} &\mathcal{L}_{kl}(\boldsymbol{x}, \boldsymbol{y}) = \\ &\mathcal{KL}[q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y}; \boldsymbol{x})||p(\boldsymbol{z}|\boldsymbol{y}; \boldsymbol{x})] + \mathcal{KL}[q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{y})||p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{y})] \\ &= \mathcal{KL}[q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y}; \boldsymbol{x})||p(\boldsymbol{z})] - \mathbb{E}_q[log\, p(\boldsymbol{y}|\boldsymbol{z}; \boldsymbol{x})] + \\ &\mathcal{KL}[q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{y})||p(\boldsymbol{z})] - \mathbb{E}_q[log\, p(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{y})] + const \\ &= tr(\lambda D \boldsymbol{\Sigma}_{z|x,y} - log \boldsymbol{\Sigma}_{z|x,y}) + \boldsymbol{\mu}_{z|x,y}^T \boldsymbol{\Lambda}_z \boldsymbol{\mu}_{z|x,y} + \\ &\frac{\gamma}{K} \sum_k (NCC(\boldsymbol{y}, \boldsymbol{x} \circ \boldsymbol{\phi}_k) + NCC(\boldsymbol{x}, \boldsymbol{y} \circ \boldsymbol{\phi}_k^{-1})) + const. \end{aligned} \quad (9)$$

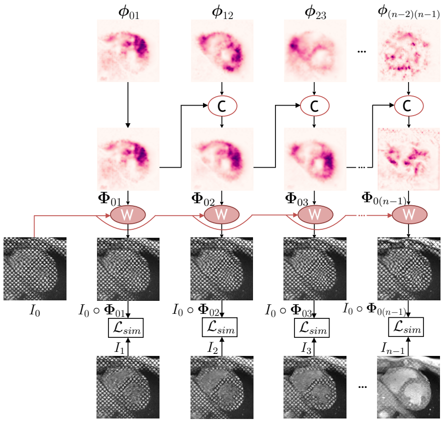

Figure 5. A composition layer C that transforms INF motion field $\boldsymbol{\phi}$ to Lagrangian motion field $\boldsymbol{\Phi}$. "W" means "warp".

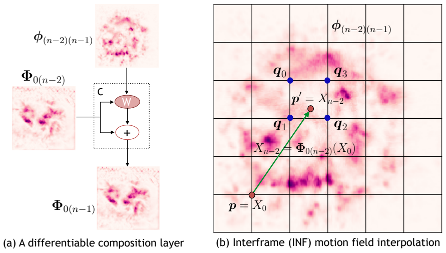

Figure 6. (a) The differentiable composition layer C. (b) INF motion field $\boldsymbol{\phi}$ interpolation at the new tracked position $\boldsymbol{p}'$.

The second term spatially smooths the mean $\boldsymbol{\mu}_{\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}}$, as we can expand it as $\boldsymbol{\mu}_{z|x,y}^T \boldsymbol{\Lambda}_z \boldsymbol{\mu}_{z|x,y} = \frac{\lambda}{2} \sum \sum_{j \in N(i)} (\boldsymbol{\mu}[i] - \boldsymbol{\mu}[j])^2$, where $N(i)$ are the neighbors of pixel i. While this is an implicit smoothness of the motion field, we also enforce the explicit smoothness of the motion field $\boldsymbol{\phi}$ by penalizing its gradients: $\mathcal{L}_{smooth}(\boldsymbol{\phi}) = \|\triangledown \boldsymbol{\phi}\|_2^2$.

Such a bi-directional registration architecture not only enforces the invertibility of the estimated motion field but also provides a path for the inverse consistency of the predicted motion field. Since the tags fade in later frames in a cardiac cycle and there exists a through-plane motion problem, we need this forward-backward constraint to obtain a more reasonable motion tracking result.

3.4. Global Lagrangian Motion Constraints

After we get all the INF motion fields in a t-MRI image sequence, we design a differentiable composition layer C to recompose them as the Lagrangian motion field $\boldsymbol{\Phi}$, as shown in Fig. 5. From Fig. 2 we can get, $\boldsymbol{\Phi}_{01} = \boldsymbol{\phi}_{01}$, $\boldsymbol{\Phi}_{0(n-1)} = \boldsymbol{\Phi}_{0(n-2)} + \boldsymbol{\phi}_{(n-2)(n-1)} (n > 2)$. However, as Fig. 6 (b) shows, the new position $\boldsymbol{p}' = X_{n-2} = \boldsymbol{\Phi}_{0(n-2)}(X_0)$ could be a sub-pixel location, and because INF motion field values are only defined at integer locations, we linearly interpolate the values between the four neighboring pixels:

$$\begin{aligned}&\boldsymbol{\phi}_{(n-2)(n-1)} \circ \boldsymbol{\Phi}_{0(n-2)}(X_0)\\ &= \sum_{\boldsymbol{q}\in N(\boldsymbol{p}')} \boldsymbol{\phi}_{(n-2)(n-1)}[\boldsymbol{q}] \prod_{d\in\{x,y\}} (1-|\boldsymbol{p}'_d - \boldsymbol{q}_d|),\end{aligned} \tag{10}$$

where $N(\boldsymbol{p}')$ are the pixel neighbors of $\boldsymbol{p}'$, and d iterates over dimensions of the motion field spatial domain. Note here we use $\boldsymbol{\phi}[\cdot]$ to denote the values of $\boldsymbol{\phi}$ at location $[\cdot]$ to differentiate it from $\boldsymbol{\phi}(\cdot)$, which means a mapping that moves one location X_{n-2} to another X_{n-1}; the same is used with $\boldsymbol{\Phi}[\cdot]$ in the following. In this formulation, we use a spatial transform layer to implement the INF motion field interpolation. Then we add the interpolated $\boldsymbol{\phi}_{(n-2)(n-1)}$ to the $\boldsymbol{\Phi}_{0(n-2)}$ and get the $\boldsymbol{\Phi}_{0(n-1)}(n > 2)$, as shown in Fig. 6 (a) (see details of computing $\boldsymbol{\Phi}$ from $\boldsymbol{\phi}$ in Algorithm 1 in Supplementary Material).

With the Lagrangian motion field $\boldsymbol{\Phi}_{0(n-1)}$, we can warp the reference frame image I_0 to any other frame at $t = n-1$: $I_0 \circ \boldsymbol{\Phi}_{0(n-1)}$. By measuring the NCC similarity between I_{n-1} and $I_0 \circ \boldsymbol{\Phi}_{0(n-1)}$, we form a global Lagrangian motion consistency constraint:

$$\mathcal{L}_g = -\sum_{n=2}^{N} NCC(I_{n-1}, I_0 \circ \boldsymbol{\Phi}_{0(n-1)}), \tag{11}$$

where N is the total frame number of a t-MRI image sequence. This global constraint is necessary to guarantee that the estimated INF motion field $\boldsymbol{\phi}$ is reasonable to satisfy a global Lagrangian motion field. Since the INF motion estimation could be erroneous, especially for large motion in between two consecutive frames, the global constraint can correct the local estimation within a much broader horizon by utilizing temporal information. Further, we also enforce the explicit smoothness of the Lagrangian motion field $\boldsymbol{\Phi}$ by penalizing its gradients: $\mathcal{L}_{smooth}(\boldsymbol{\Phi}) = \|\bigtriangledown\boldsymbol{\Phi}\|_2^2$.

To sum up, the complete loss function of our model is the weighted sum of $\mathcal{L}_{kl}$, $\mathcal{L}_{smooth}$ and $\mathcal{L}_g$:

$$\begin{aligned}\mathcal{L} = &\sum_{n=0}^{N-2}[\mathcal{L}_{kl}(I_n, I_{n+1}) + \alpha_1(\mathcal{L}_{smooth}(\boldsymbol{\phi}_{n(n+1)}) +\\ &\mathcal{L}_{smooth}(\boldsymbol{\phi}_{(n+1)n})) + \alpha_2\mathcal{L}_{smooth}(\boldsymbol{\Phi}_{0(n+1)})] + \beta\mathcal{L}_g,\end{aligned} \tag{12}$$

where α_1, α_2 and β are the weights to balance the contribution of each loss term.

4. Experiments

4.1. Dataset and Pre-Processing

To evaluate our method, we used a clinical t-MRI dataset which consists of 23 subjects' whole heart scans. Each scan set covers the 2-, 3-, 4-chamber and short-axis (SAX) views. For the SAX views, it includes several slices starting from the base to the apex of the heart ventricle; each set has approximately 10 2D slices, each of which covers a full cardiac cycle forming a 2D sequence. In total, there are 230 2D sequences in our dataset. For each sequence, the frame numbers vary from 16 $\sim$ 25. We first extracted the region of interest (ROI) from the images to cover the heart, then resampled them to the same in-plane spatial size 192×192. Each sequence was used as input to the model to track the cyclic cardiac motion. For the temporal dimension, if the frames are less than 25, we copy the last frame to fill the gap. So each input data is a 2D sequence consists of 25 frames whose spatial resolution is 192×192. We randomly split the dataset into 140, 30 and 60 sequences as the train, validation and test sets, respectively (Each set comes from different subjects). For each 2D image, we normalized the image values by first dividing them with the 2 times of median intensity value of the image and then truncating the values to be $[0, 1]$. We also did 40 times data augmentation with random rotation, translation, scaling and Gaussian noise addition.

4.2. Evaluation Metrics

Two clinical experts annotated $8 \sim 32$ landmarks on the LV MYO wall for each testing sequence, for example, as shown in Fig. 7 by the red dots; they double checked all the annotations carefully. During evaluation, we input the landmarks on the first frame and predicted their locations on the later frames by the Lagrangian motion field $\boldsymbol{\Phi}$. Following the metric used in [12], we used the root mean squared (RMS) error of distance between the centers of predicted landmark X' and ground truth landmark X to assess motion tracking accuracy. In addition, we evaluated the diffeomorphic property of the predicted INF motion field $\boldsymbol{\phi}$, using the Jacobian determinant $det(J_{\boldsymbol{\phi}}(\boldsymbol{p}))$ (detailed definitions of the two metrics in Supplementary Material).

4.3. Baseline Methods

We compared our proposed method with two conventional t-MRI motion tracking methods. The first one is HARP [37]. We reimplemented it in MATLAB (R2019a). Another one is the variational OF method[1] [11], which uses a total variation (TV) regularization term. We also compared our method with the unsupervised deep learning-based medical image registration methods VM [6] and VM-DIF [14], which are recent cutting-edge unsupervised image registration approaches. VM uses SSD (MSE) or NCC loss for training, while VM-DIF uses SSD loss. We used their official implementation code online[2], and trained VM and VM-DIF from scratch by following the optimal hyper-parameters suggested by the authors.

[1] Code is online http://www.iv.optica.csic.es/page49/page54/page54.html

[2] https://github.com/voxelmorph/voxelmorph

Method	RMS $(mm) \downarrow$	$det(J_\phi) \leqslant 0$ $(\#) \downarrow$	Time $(s) \downarrow$
HARP	3.814 ± 1.098	5950.4 ± 1709.4	124.3446 ± 21.0055
OF-TV	2.529 ± 0.726	80.3 ± 71.1	36.6764 ± 10.6163
VM (SSD)	3.799 ± 1.031	622.2 ± 390.7	$\mathbf{0.0161} \pm 0.0355$
VM (NCC)	2.856 ± 1.185	11.4 ± 12.3	0.0162 ± 0.0351
VM-DIF	3.235 ± 1.144	1.7 ± 1.6	$0.0202 \pm \mathbf{0.0332}$
Ours	$\mathbf{1.628 \pm 0.587}$	$\mathbf{0.0 \pm 0.0}$	0.0202 ± 0.0339

Table 1. Average RMS error, number of pixels with non-positive Jacobian determinant and running time.

4.4. Implementation Details

We implemented our method with Pytorch. For the FCN, the architecture is the same as in [14]. We used the Adam optimizer with a learning rate of $5e^{-4}$ to train our model. For the hyper-parameters, we set $\alpha_1 = 5$, $\alpha_2 = 1$, $\beta = 0.5$, $\gamma = -0.25$, $\lambda = 10$, via grid search. All models were trained on an NVIDIA Quadro RTX 8000 GPU. The models with the lowest loss on the validation set were selected for evaluation.

4.5. Results

4.5.1 Motion Tracking Performance

In Table 1, we show the average RMS error and the number of pixels with non-positive Jacobian determinant for baseline motion tracking methods and ours. We also show an example in Fig. 7 (full sequence results in Supplementary Material). Mean and standard deviation of the RMS errors across a cardiac cycle are shown in Fig. 8. For HARP, which is based on phase estimation, there could be missing landmark tracking results on the septal wall, due to unrealistic phase estimations, as indicated by the arrows in Fig. 7. In addition, depending on the accuracy of the phase estimation, the tracked landmarks could drift far away although the points of each landmark should be spatially close. OF-TV performs better than HARP, but it suffers from tag fading and large motion problems. The tracking results drifted a lot in the later frames. As shown in Fig. 8, the RMS error for OF-TV increased with the cardiac cycle phase. VM (NCC) is better than VM (SSD), because of the robustness of NCC loss for intensity time-variant image registration. While VM-DIF uses the SSD loss, it is better than VM (SSD) because of the diffeomorphic motion field that VM-DIF aims to learn. However, VM-DIF is worse than VM (NCC), indicating that NCC loss is more suitable for intensity time-variant image registration problems than SSD loss. VM and VM-DIF are worse than OF-TV, which suggests that we cannot apply the cutting-edge unsupervised registration methods to the t-MRI motion tracking problem without any adaptation. Our method obtains the best performance since it utilizes the NCC loss, bi-directional and global Lagrangian constraints, as well as the diffeomorphic nature of the learned motion field. The diffeomorphic attribute is also reflected by the Jacobian determinant. Our method maintains the number of pixels with non-positive Jacobian determinant as zero, which indicates the learned motion field is smooth, topology preserving and ensures one-to-one mapping.

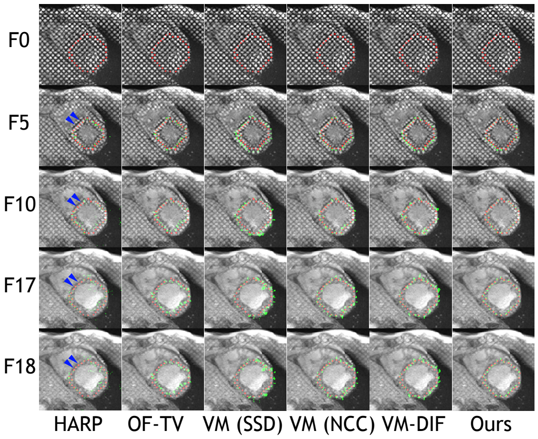

Figure 7. Motion tracking results on a t-MRI image sequence of 19 frames (best viewed zoomed in). Red is ground truth, green is prediction. “F” means “frame”.

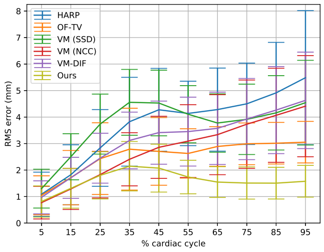

Figure 8. Mean and standard deviation of the RMS errors across a cardiac cycle for baseline methods and ours.

4.5.2 Ablation Study and Results

To compare the efficiency of tracking Lagrangian motion and INF motion, we designed two kinds of restricted models. One is to do registration between the reference and any other later frame, the other is registration between consecutive frames: A1 (forward Lagrangian tracking) and A2 (forward INF tracking). To explore the effect of bi-directional regularization, we studied the forward-backward model: A3 (A2 + backward INF tracking). We then studied the effect of explicit smoothness over the INF motion field: A4 (A3 + INF motion field ϕ smooth). To validate our proposed global Lagrangian motion constraint, we studied models with every four frames and with full sequence global constraint: A5 (A4 + every 4 frames Lagrangian constraint) and A6 (A4 + full sequence Lagrangian constraint). We also studied the effect of explicit smoothness over the La-

Model	RMS $(mm)\downarrow$	$det(J_{\Phi}) \leqslant 0\ (\#)\downarrow$
A1	2.958 ± 0.695	0.0 ± 0.0
A2	2.977 ± 1.217	0.0 ± 0.0
A3	1.644 ± 0.611	0.0 ± 0.0
A4	$1.654 \pm \mathbf{0.586}$	0.0 ± 0.0
A5	1.704 ± 0.677	0.0 ± 0.0
A6	1.641 ± 0.637	0.0 ± 0.0
Ours	$\mathbf{1.628} \pm 0.587$	$\mathbf{0.0} \pm \mathbf{0.0}$

Table 2. Ablation study results.

grangian motion field: Ours (A6 + Lagrangian motion field $\mathbf{\Phi}$ smooth).

In Table 2, we show the average RMS error and number of pixels with non-positive Jacobian determinant. We also show an example in Fig. 9 (full sequence results in Supplementary Material). The mean and standard devation of RMS errors for each model across a cardiac cycle is shown in Fig. 10. As we previously analyzed in Section 3.1, directly tracking Lagrangian motion will deduce a drifted result for large motion frames, as shown in frame $5 \sim 11$ for A1 in Fig. 9. Although forward-only INF motion tracking (A2) performs worse than A1 on average, mainly due to tag fading on later frames, bi-directional INF motion tracking (A3) is better than both A1 and A2. From Fig. 10, A3 mainly improves the performance of INF motion tracking estimation on later frames with the help of inverse consistency of the backward constraint. The explicit INF and Lagrangian motion field smoothness regularization (A4 and ours) helps to smooth the learned motion field for later frames with the prior that spatially neighboring pixels should move smoothly together. However, the smoothness constraints make it worse for the earlier (systolic) frames, which warrants a further study of a time-variant motion field smoothness constraint in the future. Our proposed global Lagrangian motion constraint greatly improved the estimation of the large INF motion (A6 and ours). As shown in Fig. 9, beginning with frame 9, the heart gets into the rapid early filling phase. INF motion in between frame 9 and 10 is so large that, without a global motion constraint (A3 and A4), the tracking results would drift a lot on the lateral wall as indicated by arrows. What's worse, such a drift error will accumulate over the following frames, which results in erroneous motion estimation on a series of frames. The proposed global constraint, however, could correct such an unreasonable INF motion estimation and a full sequence global constraint (A6) achieves better results than the segmented every 4 frames constraint (A5). All models have no non-positive Jacobian determinants, suggesting that the learned motion fields guarantee one-to-one mapping.

4.5.3 Running Time Analysis

In Table 1, we report the average inference time for motion tacking on a full t-MRI image sequence by using an Intel Xeon CPU and an NVIDIA Quadro RTX 8000 GPU for different tracking methods. While the unsupervised deep learning-based methods utilize both CPU and GPU during inference, conventional methods (HARP and OF-TV) only use the CPU. It can be noted that the learning-based method is much faster than the conventional iteration-based method. Our method can complete the inference of the full sequence in one second. In this way, we can expect very fast and accurate regional myocardial movement tracking on t-MRI images that can be used in future clinical practice.

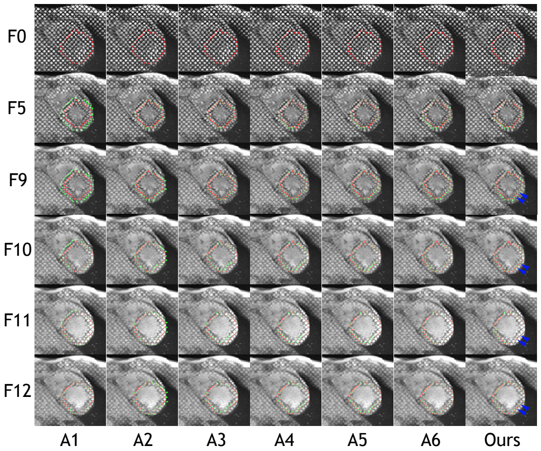

Figure 9. Ablation study results on an image sequence of 19 t-MRI frames. Red is ground truth, green is prediction.

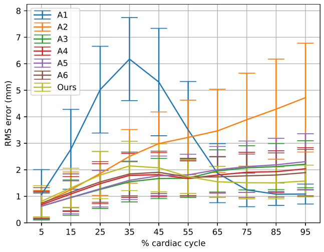

Figure 10. Mean and standard deviation of the RMS error during the entire cardiac cycle for the ablation study models and ours.

5. Conclusions

In this work, we proposed a novel bi-directional unsupervised diffeomorphic registration network to track regional myocardium motion on t-MRI images. We decomposed the Lagrangian motion tracking into a sequence of INF motion tracking, and used global constraints to correct unreasonable INF motion estimation. Experimental results on the clinical t-MRI dataset verified the effectiveness and efficiency of the proposed method.

References

[1] Mihaela Silvia Amzulescu, M De Craene, H Langet, Agnes Pasquet, David Vancraeynest, Anne-Catherine Pouleur, Jean-Louis Vanoverschelde, and BL Gerber. Myocardial strain imaging: review of general principles, validation, and sources of discrepancies. *European Heart Journal-Cardiovascular Imaging*, 20(6):605–619, 2019. 1

[2] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 924–931. Springer, 2006. 4

[3] John Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, 2007. 4

[4] Brian B Avants, Nicholas J Tustison, Gang Song, Philip A Cook, Arno Klein, and James C Gee. A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, 2011. 5

[5] Leon Axel and Lawrence Dougherty. Mr imaging of motion with spatial modulation of magnetization. *Radiology*, 171(3):841–845, 1989. 2

[6] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9252–9260, 2018. 3, 6

[7] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *European conference on computer vision*, pages 25–36. Springer, 2004. 2

[8] Thomas Brox and Jitendra Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE transactions on pattern analysis and machine intelligence*, 33(3):500–513, 2010. 2

[9] Xiaohuan Cao, Jianhua Yang, Jun Zhang, Dong Nie, Minjeong Kim, Qian Wang, and Dinggang Shen. Deformable image registration based on similarity-steered cnn regression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 300–308. Springer, 2017. 3

[10] Xiaohuan Cao, Jianhua Yang, Jun Zhang, Qian Wang, Pew-Thian Yap, and Dinggang Shen. Deformable image registration using a cue-aware deep regression network. *IEEE Transactions on Biomedical Engineering*, 65(9):1900–1911, 2018. 3

[11] Noemi Carranza-Herrezuelo, Ana Bajo, Filip Sroubek, Cristina Santamarta, Gabriel Cristóbal, Andrés Santos, and María J Ledesma-Carbayo. Motion estimation of tagged cardiac magnetic resonance images using variational techniques. *Computerized Medical Imaging and Graphics*, 34(6):514–522, 2010. 2, 6

[12] Raghavendra Chandrashekara, Raad H Mohiaddin, and Daniel Rueckert. Analysis of 3-d myocardial motion in tagged mr images using nonrigid image registration. *IEEE Transactions on Medical Imaging*, 23(10):1245–1250, 2004. 3, 6

[13] Ting Chen, Xiaoxu Wang, Sohae Chung, Dimitris Metaxas, and Leon Axel. Automated 3d motion tracking using gabor filter bank, robust point matching, and deformable models. *IEEE Transactions on Medical Imaging*, 29(1):1–11, 2009. 2

[14] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 729–738. Springer, 2018. 3, 4, 5, 6, 7

[15] Bob D de Vos, Floris F Berendsen, Max A Viergever, Marius Staring, and Ivana Išgum. 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*, pages 204–212. Springer, 2017. 3

[16] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 3

[17] Safaa M ElDeeb and Ahmed S Fahmy. Accurate harmonic phase tracking of tagged mri using locally-uniform myocardium displacement constraint. *Medical Engineering & Physics*, 38(11):1305–1313, 2016. 2

[18] Berthold KP Horn and Brian G Schunck. Determining optical flow. In *Techniques and Applications of Image Understanding*, volume 281, pages 319–331. International Society for Optics and Photonics, 1981. 2

[19] Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5754–5763, 2019. 3

[20] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 3

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[22] Julian Krebs, Hervé Delingette, Boris Mailhé, Nicholas Ayache, and Tommaso Mansi. Learning a probabilistic model for diffeomorphic registration. *IEEE transactions on medical imaging*, 38(9):2165–2176, 2019. 1, 3

[23] Julian Krebs, Tommaso Mansi, Hervé Delingette, Li Zhang, Florin C Ghesu, Shun Miao, Andreas K Maier, Nicholas Ayache, Rui Liao, and Ali Kamen. Robust non-rigid registration through agent-based action learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 344–352. Springer, 2017. 3

[24] Hsueh-Ying Lai, Yi-Hsuan Tsai, and Wei-Chen Chiu. Bridging stereo matching and optical flow via spatiotemporal correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1890–1899, 2019. 3

[25] Maria J Ledesma-Carbayo, J Andrew Derbyshire, Smita Sampath, Andrés Santos, Manuel Desco, and Elliot R

McVeigh. Unsupervised estimation of myocardial displacement from tagged mr sequences using nonrigid registration. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 59(1):181–189, 2008. 3

[26] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selflow: Self-supervised learning of optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4571–4580, 2019. 3

[27] Xiaofeng Liu, Khaled Z Abd-Elmoniem, Maureen Stone, Emi Z Murano, Jiachen Zhuo, Rao P Gullapalli, and Jerry L Prince. Incompressible deformation estimation algorithm (idea) from tagged mr images. *IEEE transactions on medical imaging*, 31(2):326–340, 2011. 2

[28] Xiaofeng Liu and Jerry L Prince. Shortest path refinement for motion estimation from tagged mr images. *IEEE Transactions on Medical Imaging*, 29(8):1560–1572, 2010. 2

[29] Marco Lorenzi, Nicholas Ayache, Giovanni B Frisoni, Xavier Pennec, Alzheimer's Disease Neuroimaging Initiative (ADNI, et al. Lcc-demons: a robust and accurate symmetric diffeomorphic registration algorithm. *NeuroImage*, 81:470–483, 2013. 5

[30] Kristin McLeod, Adityo Prakosa, Tommaso Mansi, Maxime Sermesant, and Xavier Pennec. An incompressible log-domain demons algorithm for tracking heart tissue. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 55–67. Springer, 2011. 3

[31] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. *arXiv preprint arXiv:1711.07837*, 2017. 3

[32] Etienne Mémin and Patrick Pérez. Dense estimation and object-based segmentation of the optical flow with robust techniques. *IEEE Transactions on Image Processing*, 7(5):703–719, 1998. 2

[33] Tony CW Mok and Albert Chung. Fast symmetric diffeomorphic image registration with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4644–4653, 2020. 3, 4

[34] Pedro Morais, Brecht Heyde, Daniel Barbosa, Sandro Queirós, Piet Claus, and Jan D'hooge. Cardiac motion and deformation estimation from tagged mri sequences using a temporal coherent image registration framework. In *International Conference on Functional Imaging and Modeling of the Heart*, pages 316–324. Springer, 2013. 3

[35] Manuel A Morales, David Izquierdo-Garcia, Iman Aganj, Jayashree Kalpathy-Cramer, Bruce R Rosen, and Ciprian Catana. Implementation and validation of a three-dimensional cardiac motion estimation network. *Radiology: Artificial Intelligence*, 1(4):e180080, 2019. 1

[36] Marc Niethammer, Roland Kwitt, and Francois-Xavier Vialard. Metric learning for image registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8463–8472, 2019. 3

[37] Nael F Osman, William S Kerwin, Elliot R McVeigh, and Jerry L Prince. Cardiac motion tracking using cine harmonic phase (harp) magnetic resonance imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 42(6):1048–1060, 1999. 2, 6

[38] Nael F Osman, Elliot R McVeigh, and Jerry L Prince. Imaging heart motion using harmonic phase mri. *IEEE transactions on medical imaging*, 19(3):186–202, 2000. 2

[39] Zhen Qian, Qingshan Liu, Dimitris N Metaxas, and Leon Axel. Identifying regional cardiac abnormalities from myocardial strains using nontracking-based strain estimation and spatio-temporal tensor analysis. *IEEE Transactions on Medical Imaging*, 30(12):2017–2029, 2011. 2

[40] Chen Qin, Wenjia Bai, Jo Schlemper, Steffen E Petersen, Stefan K Piechnik, Stefan Neubauer, and Daniel Rueckert. Joint learning of motion estimation and segmentation for cardiac mr image sequences. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 472–480. Springer, 2018. 1

[41] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 12240–12249, 2019. 3

[42] Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec. Svf-net: Learning deformable image registration using shape matching. In *International conference on medical image computing and computer-assisted intervention*, pages 266–274. Springer, 2017. 3

[43] Nicolas Rougon, Caroline Petitjean, Françoise Prêteux, Philippe Cluzel, and Philippe Grenier. A non-rigid registration approach for quantifying myocardial contraction in tagged mri using generalized information measures. *Medical Image Analysis*, 9(4):353–375, 2005. 3

[44] Zhengyang Shen, Xu Han, Zhenlin Xu, and Marc Niethammer. Networks for joint affine and non-parametric image registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4224–4233, 2019. 3

[45] Zhengyang Shen, François-Xavier Vialard, and Marc Niethammer. Region-specific diffeomorphic metric mapping. In *Advances in Neural Information Processing Systems*, pages 1098–1108, 2019. 3

[46] Wenzhe Shi, Xiahai Zhuang, Haiyan Wang, Simon Duckett, Duy VN Luong, Catalina Tobon-Gomez, KaiPin Tung, Philip J Edwards, Kawal S Rhode, Reza S Razavi, et al. A comprehensive cardiac motion estimation framework using both untagged and 3-d tagged mr images based on nonrigid registration. *IEEE transactions on medical imaging*, 31(6):1263–1275, 2012. 3

[47] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 3

[48] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. *arXiv preprint arXiv:2003.12039*, 2020. 3

[49] Liang Wang, Patrick Clarysse, Zhengjun Liu, Bin Gao, Wanyu Liu, Pierre Croisille, and Philippe Delachartre. A gradient-based optical-flow cardiac motion estimation method for cine and tagged mr images. *Medical image analysis*, 57:136–148, 2019. 2

[50] Xiaoxu Wang, Dimitis Metaxas, Ting Chen, and Leon Axel. Meshless deformable models for lv motion analysis. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[51] Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, and Wei Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8071–8081, 2019. 3

[52] Andreas Wedel, Daniel Cremers, Thomas Pock, and Horst Bischof. Structure-and motion-adaptive regularization for high accuracy optic flow. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1663–1668. IEEE, 2009. 2

[53] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1983–1992, 2018. 3

[54] Hanchao Yu, Xiao Chen, Humphrey Shi, Terrence Chen, Thomas S Huang, and Shanhui Sun. Motion pyramid networks for accurate and efficient cardiac motion estimation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*, pages 436–446. Springer, 2020. 1

[55] Hanchao Yu, Shanhui Sun, Haichao Yu, Xiao Chen, Honghui Shi, Thomas S Huang, and Terrence Chen. Foal: Fast online adaptive learning for cardiac motion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4313–4323, 2020. 1

[56] Shengyu Zhao, Yue Dong, Eric I Chang, Yan Xu, et al. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10600–10610, 2019. 3

[57] Qiao Zheng, Hervé Delingette, and Nicholas Ayache. Explainable cardiac pathology classification on cine mri with motion characterization by semi-supervised learning of apparent flow. *Medical image analysis*, 56:80–95, 2019. 1