Medical Image Analysis 54 (2019) 20–29

Contents lists available at ScienceDirect

Medical Image Analysis

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

An image interpolation approach for acquisition time reduction in navigator-based 4D MRI

Neerav Karani[1,*], Lin Zhang[1], Christine Tanner, Ender Konukoglu

Biomedical Image Computing Group, ETH Zurich, Switzerland

ARTICLE INFO

Article history:
Received 15 August 2018
Revised 1 February 2019
Accepted 12 February 2019
Available online 13 February 2019

Keywords:
Temporal image interpolation
Motion field estimation

ABSTRACT

Navigated 2D multi-slice dynamic Magnetic Resonance (MR) imaging enables high contrast 4D MR imaging during free breathing and provides in-vivo observations for treatment planning and guidance. Navigator slices are vital for retrospective stacking of 2D data slices in this method. However, they also prolong the acquisition sessions. Temporal interpolation of navigator slices can be used to reduce the number of navigator acquisitions without degrading specificity in stacking. In this work, we propose a convolutional neural network (CNN) based method for temporal interpolation, with motion field prediction as an intermediate step. The proposed formulation incorporates the prior knowledge that a motion field underlies changes in the image intensities over time. Previous approaches that interpolate directly in the intensity space are prone to produce blurry images or even remove structures in the images. Our method avoids such problems and faithfully preserves the information in the image. Further, an important advantage of our formulation is that it provides an unsupervised estimation of bi-directional motion fields. These motion fields can potentially be used to halve the number of registrations required during 4D reconstruction, thus substantially reducing the reconstruction time. These advantages are achieved while preserving 4D reconstruction quality as compared to that with the true navigators.

1. Introduction

Involuntary motion of anatomical structures due to factors such as breathing, peristalsis and heart beat is an important concern in image-guided therapy applications, such as planning and guiding radiotherapy (Bert and Durante, 2011) and high intensity focused ultrasound therapy (Arnold et al., 2011). For instance, if such motion is not taken into account in radiotherapy, it may lead to dose distribution degradation (Lambert et al., 2005), reducing the efficacy of the treatment and irradiating healthy tissue. Visualizing and quantifying these motion patterns is thus an essential component in these applications and is enabled by recent advances in dynamic volumetric (4D) imaging (Li et al., 2008).

4D imaging can be done with either CT (Low et al., 2003; Keall et al., 2004) or MRI (Cai et al., 2011; Von Siebenthal et al., 2007). The benefits of the latter over the former include better soft-tissue contrast, flexibility in choosing plane orientations as well as not exposing patients to ionizing radiation, which allows acquiring long sequences to capture motion irregularities. However, due to inherent trade-offs in MR imaging between image resolution, acquisition time and signal-to-noise ratio, 4D MRIs cannot be acquired by imaging entire volumes of interest at each instant in time. Consequently, several techniques have been proposed to overcome these trade-offs. Among these techniques, the ones that allow for free-breathing imaging typically use the following strategy: acquire 2D image slices from the volume of interest across several breathing cycles and retrospectively create the 4D image by stacking together 2D slices acquired at the same phase in different breathing cycles. Such *slice-stacking* techniques (Von Siebenthal et al., 2007; Tryggestad et al., 2013; Baumgartner et al., 2013) differ from one another in the metric they employ to identify temporally corresponding slices from different breathing cycles. In this article, we focus on *navigated 2D multi-slice imaging* (Von Siebenthal et al., 2007), a particular variant among slice-stacking techniques for 4D MRI.

In *navigated 2D multi-slice imaging* (Von Siebenthal et al., 2007), so-called *navigator* slices or *navigators* are acquired, interleaved with the 2D data slice acquisitions. The navigators are dedicated acquisitions, whose purpose is to tag their temporally adjacent data slices to phases of the breathing cycle. The scheme acquires navigator slices N_t (at the same anatomical location) and data slices D_t^p (at different locations p to cover the volume of interest) alternately over several breathing cycles, as shown in the illustration in Fig. 1(a). Here, t denotes the temporal index of the N or

[*] Corresponding author.
E-mail address: nkarani@vision.ee.ethz.ch (N. Karani).
[1] Both authors have contributed equally to this manuscript.

https://doi.org/10.1016/j.media.2019.02.008
1361-8415/© 2019 Elsevier B.V. All rights reserved.

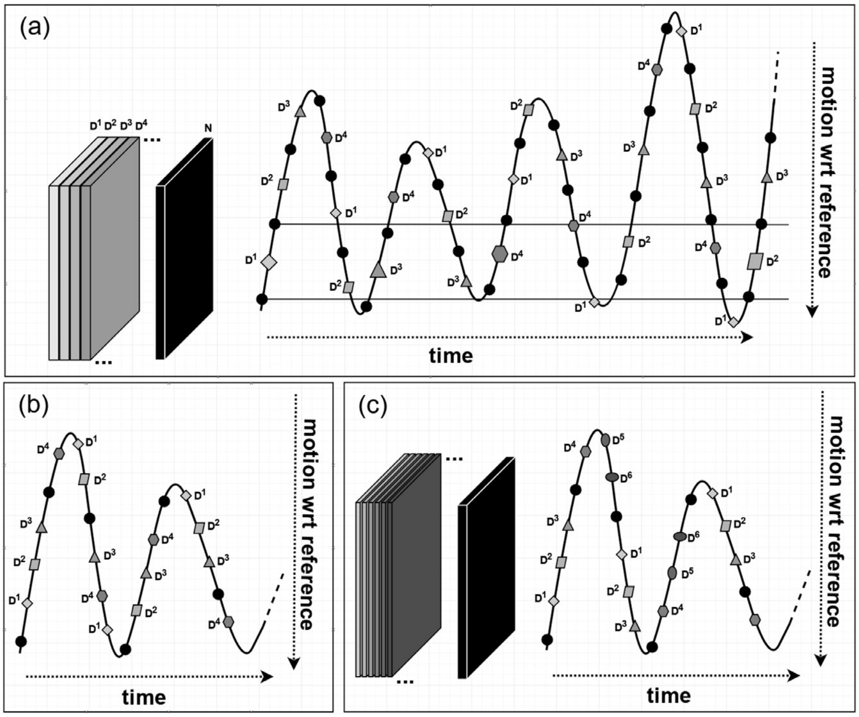

Fig. 1. Acquisition schemes of navigated 2D multi-slice imaging showing (a) conventional approach (Von Siebenthal et al., 2007) interleaving 1 navigator slice *N* and 1 data slice D^p, (b-c) modified approaches interleaving 1 *N* and 3 D^p, which can be used (b) to reduce acquisition time with the same image resolution as before or (c) to improve image resolution by acquiring more data slices to cover the same volume of interest. The black circles denote the navigator acquisitions, whereas the different shades of gray and geometric shapes refer to different data slices. The horizontal lines in (a) depict the 3D volume reconstruction procedure for the first shown data slice. The enlarged data slices at other depths are the ones that are selected to form the 3D volume, by comparing the corresponding navigator positions.

D acquisition in the sequence. After the acquisition is complete, each navigator is registered to a selected end-exhalation navigator, which we call as the reference navigator. The mean 2D motion of the resultant motion field over the region of interest is computed as an approximate descriptor of the breathing phase. Subsequently, for each navigator, data slices enclosed by other navigators acquired at the most similar phase of the breathing cycle are stacked together to form a 3D image at that phase. Combining volumes corresponding to each acquired navigator provides the final 4D reconstruction over several breathing cycles. Note that there are two main assumptions in this 4D MRI technique. Firstly, as the data slice sorting is based on the motion in the selected region of the navigators, consistent reconstruction for moving organs will only be achieved if their motion is correlated with the motion in the selected region. Secondly, out-of-plane motion in the navigator slices is assumed to be minimal, so that the registration, which measures only in-plane motion, provides an accurate descriptor of the position in the breathing cycle. These assumptions are, however, easily met in practice due to the orientation selection flexibility in MRI and by choosing a navigator location through the organ of interest, which shows stable structures during free-breathing.

Other slice-stacking techniques without navigators has been proposed by using external breathing signal (Tryggestad et al., 2013) or consistency between adjacent data slices (Baumgartner et al., 2013). However, navigators enable continuous organ motion quantification, which might not be externally measurable (e.g. drift of the liver (Von Siebenthal et al., 2007)) or hard to accurately estimate from the data slices, and hence potentially provide superior reconstructions. An additional benefit of this technique is that it is especially useful for capturing irregular breathing patterns as it can provide long and uninterrupted observations during free-breathing.

While navigators are essential for the specificity in slice stacking in navigated multi-slice imaging, they also prolong the acquisition time. If it were possible to retain the benefits of dedicated navigators with relatively infrequent navigator acquisitions, the saved time could be utilized to either reduce the overall acquisition time or to improve through-plane resolution, using the same time to acquire more data slices. These two scenarios are depicted in the modified acquisition schemes shown in Fig. 1(b) and (c) respectively. This motivates the idea of acquiring fewer navigators and temporally interpolating these to predict the missing ones.

With this motivation, (Karani et al., 2017) proposed a convolutional neural network (CNN) based approach for temporal interpolation of navigators. Their CNN takes as inputs a fixed number of acquired images and learns to predict the missing images directly in the intensity space. This approach, which we call the *Simple Convolutional Interpolation Network* (SCIN), is a 'black-box' formulation that does not incorporate any prior information about the interpolation process. Image prediction is guided only by the cost function used to optimize the network parameters. The issue with this is that it is unclear whether the image similarity measures that are generally used as cost functions suffice to ensure fidelity of the generated images to the original images. Indeed, Fig. 2b shows a case where an image interpolated using SCIN is quite blurry and misses several liver and lung structures present in the original image.

In this article, we propose an interpolation method that incorporates the prior knowledge that a motion field underlies the difference between images acquired at different times. We note that in scenarios where an image sequence captures anatomical structures in motion (without induced contrast changes), the content of the images remains largely unchanged over time and issues such as occlusion are not pertinent. Further, if the principal direction of motion is in the plane of the 2D images, the chances of structures going out of the image or new structures coming into the image due to off-plane motion may be minimal. Under these assumptions, each image can be viewed as a spatially transformed

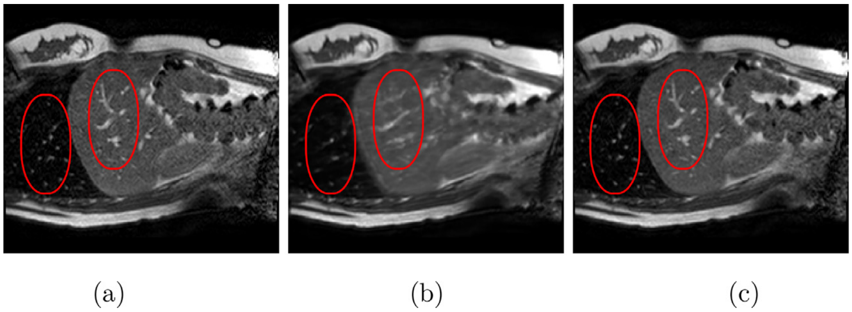

Fig. 2. (a) Ground truth and (b,c) interpolated images from (b) baseline (SCIN) and (c) proposed method (MFIN). The image interpolated via SCIN is heavily blurred and misses several lung and liver structures, while the proposed method is able to preserve the details in the ground truth image.

version of its temporal neighbours. Indeed, the retrospective sorting based on navigator slices in the 4D reconstruction procedure of (Von Siebenthal et al., 2007) itself relies on the same assumptions. This observation leads us to incorporate motion field prediction as an intermediate step for the interpolation problem, which removes the ability of the CNN to directly change image intensities and enables the regularization of the predicted motion fields. We hypothesize that this formulation makes changes in image structure unlikely, leading to more plausible predicted images. We train a CNN to take as input several acquired navigators and predict the motion fields between the image to be interpolated and its known two neighbours. Any of these two motion fields can then be used to wrap the corresponding known neighbouring image to obtain the missing image. We call this network the *Motion Field Interpolation Network* (MFIN). MFIN is trained end-to-end using only navigator images, without ground truth motion fields. Indeed, an important advantage of our interpolation formulation is that it provides an unsupervised estimation of the motion fields. In the particular setting of navigated multi-slice imaging like (Von Siebenthal et al., 2007), each navigator has to be registered to a reference image during the retrospective 4D reconstruction procedure. The motion fields obtained by our interpolation framework can be potentially used to halve the number of these registrations, substantially reducing the computational effort of 4D reconstruction.

A preliminary version of this study has been presented at a conference (Zhang et al., 2018). In that paper, we introduced the idea of temporal interpolation via motion field prediction and a novel cyclic consistency training loss to regularize the predicted motion fields. The proposed methods were extensively evaluated with respect to several intensity-based and registration-based metrics. In this work, we extend (Zhang et al., 2018) in the following ways. As previously noted, we believe that it is an important benefit of our method that it is trained end-to-end without requiring ground truth motion fields. Here, we also consider a scenario where our interpolation network is trained in a supervised manner, by providing 'ground-truth' motion fields obtained from a gold-standard registration algorithm. The aim of this experiment is to examine if the unsupervised case can already achieve the performance obtained from the supervised training. Further, we carry out an ablation study to investigate the benefit of estimating bi-directional motion fields, as the motion field in either one of the two directions is sufficient for interpolation. Finally, we evaluate the effect of the proposed interpolation method on the 4D reconstructions obtained from the interpolated navigators. Our results indicate that the quality of the 4D reconstructions is well preserved even with the interpolated navigators.

2. Related work

Temporal image interpolation in the medical imaging context has been mainly suggested for ultrasound imaging (Lee et al., 2003; Nam et al., 2006; Zhang et al., 2011). These methods explicitly track pixel-wise correspondences between neighbouring images via optical flow estimation or non-linear registration. An advantage of such methods is that they typically estimate the underlying motion fields as part of the interpolation. Yet, they often make simplistic assumptions regarding the shape and dynamics of the motion trajectory such as linear, constant velocity. With the recent surge of methods based on neural networks, *end-to-end learning*-based solutions directly predict in-between images given surrounding ones, skipping the motion field estimation. In this line, Karani et al. (2017) proposed the aforementioned SCIN for interpolating navigators for 4D-MRI reconstruction. In computer vision, variants of CNNs have been suggested for interpolation (Long et al., 2016) and video frame prediction (Srivastava et al., 2015; Goroshin et al., 2015; Mathieu et al., 2015). A common feature of these methods is that the image predictions are made directly in the intensity space.

Prediction of image intensities from scratch may be difficult, leading to blurry results, or even distortion of image structures. To tackle this, some works have suggested using the content of the known images. Yeh et al. (2016) propose to use a variational autoencoder (VAE) (Kingma and Welling, 2013) to learn bidirectional motion fields between two known images. Then, without explicit training for interpolation, motion fields from the known images to an in-between image are predicted by linear interpolation in the latent space of the VAE. This approach produces sharp interpolated facial expression images, but it is unclear whether it would be able to faithfully interpolate breathing motion patterns in abdominal navigators, which also requires the prediction of the linear interpolation coefficient. In (Jiang et al., 2017), a CNN directly predicts bi-directional motion fields between the known images, which are further refined using another CNN to account for occlusions and then combined to produce the interpolated image. In (Niklaus et al., 2017), interpolation is formulated as a local convolution over the known images and a CNN predicts a convolutional kernel for each pixel of the interpolated image. Although these methods incorporate prior information about the image content into the interpolation framework, they do not readily provide the motion fields from the known images to the final predicted image. Note that in both (Yeh et al., 2016; Jiang et al., 2017), bi-directional motion fields are combined to obtain the interpolated image, as the emphasis is on dealing with interpolation scenarios including occlusions. Such combination renders the methods unable to provide the final motion field between the predicted and the known images.

The underlying motion estimation problem has been separately studied, either requiring ground truth flow fields for training (Fischer et al., 2015; Ilg et al., 2017; Ranjan and Black, 2017) or in an unsupervised fashion by relying on reconstruction of warped images using predicted flow fields (Patraucean et al., 2015; Jason et al., 2016). The relationship between the problems of interpolation and motion field estimation has been exploited in (Long et al., 2016) to find dense correspondences between input images via saliency maps (Simonyan et al., 2013) of an interpolation CNN. Instead of this indirect approach of using interpolation to find motion fields, MFIN takes the direct route and interpolates by first estimating the underlying motion.

Another related problem is that of image registration. CNNs have been proposed for learning image registration using known ground truth motion (Dosovitskiy et al., 2015) or gold-standard registration results (Yang et al., 2017), or in an unsupervised manner within an optimization framework (de Vos et al., 2017). While the registration problem is one of motion field estimation between two known images, the interpolation problem is to predict missing images and here, we are additionally interested in estimating the underlying motion.

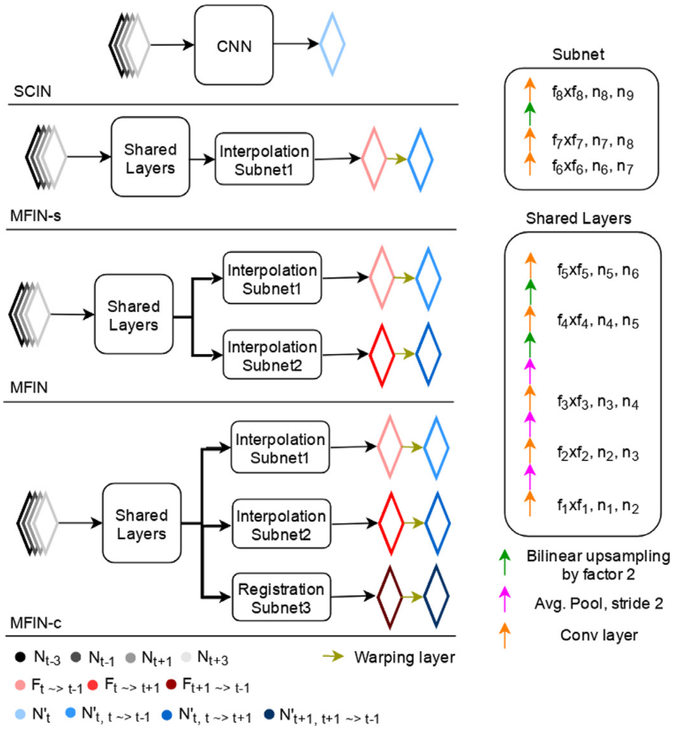

Fig. 3. Architectures of the Simple Convolutional Interpolation Network (SCIN) (Karani et al., 2017), the proposed Motion Field Interpolation Network (MFIN), its reduction to a single field prediction (MFIN-s) and its extension for incorporating cyclic consistency (MFIN-c). For the convolutional layers, f_i, n_i and n_{i+1} indicate the filter size and number of input and output channels respectively in the *i*th layer.

3. Methods

As described in Section 1, the overall acquisition scheme in navigated multi-slice imaging is $\{N_1, D_1^1, N_2, D_2^2, N_3, D_3^3, N_4, \ldots\}$, where the subscript for the navigator or the data slices represents their temporal index in the acquisition sequence, while the superscript for the data slices refers to their anatomical location. For the rest of the article, unless mentioned otherwise, we consider only the navigator sequence (i.e. $\{N_t\}$, where $t = 1, 2, 3, 4, \ldots$) as we are interested in interpolation therein. We consider a scenario where the temporal resolution of the acquired navigator sequence is sought to be doubled. That is, missing navigators $\{N_4, N_6, N_8, \ldots\}$ are to be interpolated using the acquired navigators $\{N_1, N_3, N_5, N_7, \ldots\}$. Following (Karani et al., 2017), where temporal context beyond immediate neighbours has been shown to be important for interpolating amidst non-linear motion, we provide two images each from the past and the future as inputs to MFIN. Thus, in order to interpolate N_t, the inputs to the network are N_{t-3}, N_{t-1}, N_{t+1} and N_{t+3}.

The general architecture of MFIN is shown in Fig. 3. The input images pass through shared convolutional layers before diverging into two sub-networks. Each sub-network predicts the motion field from the image to be interpolated, N_t, to one of its neighbours (N_{t-1} or N_{t+1}). The motion field predicted by each sub-network ($\mathbf{F}_{t\to t-1}$ or $\mathbf{F}_{t\to t+1}$) is used to warp the corresponding neighbouring image using bilinear interpolation to predict N_t independently. The warping layer is described in detail in Section 3.2. The loss function used to optimize the network parameters (discussed in Section 3.1) is defined to measure the dissimilarity of the interpolated and the ground truth images, thus not requiring ground truth motion fields. While either one of the two sub-networks is sufficient for interpolation, we still predict the displacement fields in both directions in view of potential inductive bias promoted by multi-task learning (Caruana, 1998). To investigate if the bidirectional field prediction is indeed beneficial or not, we train an ablated network MFIN-s(ingle) (see Fig. 3), where the motion field is predicted only in one direction. Additionally, in an extension to our base model called MFIN-c(ycle) described in Section 3.3, we utilize the bidirectional motion fields to enforce a cyclic consistency constraint.

3.1. Loss functions

The loss function for MFIN, shown in Eq. (1), consists of a reconstruction loss term (L_{recon}) and a regularization term (L_{reg}). L_{recon} (Eq. (2)) is the sum of the reconstruction errors from the two sub-networks, where $N'_{t,s}$ denotes the prediction for image N_t by warping the image N_s according to the estimated motion field $\mathbf{F}_{t\to s}$, i.e. defined by displacement vectors originating at pixel locations in N_t pointing to N_s. The form of the reconstruction loss must capture the desired notion of image dissimilarity between the predicted and the ground truth image. The mean-squared-error in intensity space, which is generally used as the reconstruction loss, might not be robust to intensity scaling. Such scaling might be relevant due to saturation effects in the employed interleaved MR acquisition, where the navigator image is saturated when the preceding data slice is at a similar location. Since the MFIN formulation simply moves the pixel-intensities from one of the known images to obtain the missing image, it cannot account for such effects. Importantly, such intensity differences are not crucial for the application of 4D reconstruction as long as the estimated motion fields are accurate. With this motivation, we investigate the use of the structural similarity (SSIM) index (Wang et al., 2004) as a similarity measure in addition to the generally used L_2 intensity loss. SSIM between two image patches p_t and p_s (P_{SSIM}) is defined in Eq. (5), where μ_t, μ_s are patch means, σ_t^2, σ_s^2 are patch variances, σ_{ts} is the covariance of the two patches and c_1, c_2 are constants. SSIM between entire images is based on the mean of SSIM between corresponding patches, and turned into a dissimilarity (i.e. loss L_{SSIM}) by taking the negative value, see Eq. (4). SSIM takes into account local correlation between image patches and therefore, may be more robust to the aforementioned intensity scalings. It has also been shown to preserve low-level structure (Snell et al., 2017). Finally, it is differentiable and can thus be readily integrated into gradient based optimization. For L_{reg}, we employ total variation regularization (Eq. (3)) to promote smoothness in the predicted motion fields, while allowing for sharp gradients to cope with sliding interfaces.

$$L_{\text{total,MFIN}} = L_{\text{recon,MFIN}} + \lambda_1 L_{\text{reg,MFIN}} \tag{1}$$

$$L_{\text{recon,MFIN}} = L\big(N_t, N'_{t,t-1}\big) + L\big(N_t, N'_{t,t+1}\big) \text{ with } L \in \{L_2, L_{\text{SSIM}}\} \tag{2}$$

$$L_{\text{reg,MFIN}} = ||\nabla \mathbf{F}_{t\to t-1}||_1 + ||\nabla \mathbf{F}_{t\to t+1}||_1 \tag{3}$$

$$L_{\text{SSIM}}(N_t, N_s) = -\frac{1}{n_P} \sum_{p_t \in N_t,\, p_s \in N_s}^{n_P} P_{\text{SSIM}}(p_t, p_s) \tag{4}$$

$$P_{\text{SSIM}}(p_t, p_s) = \frac{(2\mu_t\mu_s + c_1)(2\sigma_{ts} + c_2)}{(\mu_t^2 + \mu_s^2 + c_1)(\sigma_t^2 + \sigma_s^2 + c_2)} \tag{5}$$

3.2. Differentiable warping

Differentiable warping / sampling for target image generation by applying a flow field to a source image was introduced in (Jaderberg et al., 2015) and subsequently employed in several image synthesis works (Zhou et al., 2016; Liu et al., 2017; Jiang et al., 2017). It works as follows. Motion field $\mathbf{F}_{t\to s}$ predicted by the network contains, for each pixel location i in the target image N_t, the relative sampling location in the source image N_s. Starting with the image pixel position grid $\mathbf{G}$ of N_t and N_s, and using motion field $\mathbf{F}_{t\to s}$, we obtain the warped image grid $\mathbf{G}_{t,s} = \mathbf{G} + \mathbf{F}_{t\to s}$, where each element $\mathbf{G}_{t,s}(i) = (x(i), y(i))$ is the absolute 2D coordinate location in the source image N_s to be used for constructing the warped image $N'_{t,s}$. As these locations may not necessarily coincide with pixel centers $\mathbf{G}$ in N_s, we employ bilinear interpolation of the 4 closest neighbours in $\mathbf{G}$ (denoted as neighbourhood $\mathcal{N}_4(\mathbf{G}_{t,s}(i))$) to obtain the prediction (see Eq. (6)).

$$N'_{t,s}(i) = \sum_{j \in \mathcal{N}_4(\mathbf{G}_{t,s}(i))} N_s(j)\, \Big(1 - |x(i) - x(j)|\Big)\, \Big(1 - |y(i) - y(j)|\Big) \tag{6}$$

3.3. Cycle consistency

Cycle consistency has been shown to be an effective regularizer in registration problems (Christensen and Johnson, 2001; Gass et al., 2015) as well as in other contexts such as image generation in deep neural networks (Zhu et al., 2017). The MFIN architecture can be readily extended to include such cyclic consistency between estimated motion fields. For this, we add another sub-network to the MFIN for estimating the motion field between the two, always known, neighbours of the image to be interpolated, i.e. $\mathbf{F}_{t+1\to t-1}$. We denote the extended network by MFIN-c(ycle) and optimize for the registration and two interpolation tasks jointly by minimizing the cost function in Eq. (7). The reconstruction (Eq. (8)) and regularization (Eq. (9)) terms in the loss function are extended to include the extra sub-network. Finally, Eq. (10) shows the cycle consistency loss term, where $\odot$ denotes the pixel-wise composition of the transformations, i.e. $\mathbf{F}_{t+1\to t-1}(\mathbf{F}_{t\to t+1})$, achieved via bilinear interpolation of $\mathbf{F}_{t+1\to t-1}$. We hypothesize that the registration sub-network has an easier task as it has to estimate the motion field between two known images, while the interpolation sub-networks have to predict the missing image in addition to estimating the corresponding motion fields. Thus, $\mathbf{F}_{t+1\to t-1}$ may be more accurate than $\mathbf{F}_{t\to t-1}$, $\mathbf{F}_{t\to t+1}$ and enforcing cycle consistency may help to correct errors in the latter two.

$$L_{\text{total,MFIN-c}} = L_{\text{recon,MFIN-c}} + \lambda_1 L_{\text{reg,MFIN-c}} + \lambda_2 L_{\text{cycle,MFIN-c}} \tag{7}$$

$$L_{\text{recon,MFIN-c}} = L\Big(N_t, N'_{t,t-1}\Big) + L\Big(N_t, N'_{t,t+1}\Big) + L\Big(N_{t-1}, N'_{t-1,t+1}\Big) \tag{8}$$

$$L_{\text{reg,MFIN-c}} = ||\nabla \mathbf{F}_{t\to t-1}||_1 + ||\nabla \mathbf{F}_{t\to t+1}||_1 + ||\nabla \mathbf{F}_{t+1\to t-1}||_1 \tag{9}$$

$$L_{\text{cycle,MFIN-c}} = ||\mathbf{F}_{t\to t-1} - (\mathbf{F}_{t+1\to t-1} \odot \mathbf{F}_{t\to t+1})||_2 \tag{10}$$

3.4. Supervised training with ground truth motion fields

So far, we have described so-called *end-to-end* training methods for the interpolation networks. That is, the training loss function was defined on the interpolated image obtained by warping a neighbouring image with the predicted motion field, with only regularization losses applied directly on the motion fields. As ground truth motion fields are not required in this formulation and the true navigators that drive the loss function are obtained *for free* from the fully-acquired navigator sequences, we term these training methods as *unsupervised* interpolation methods. To investigate the effect of *supervised* training on the interpolation accuracy, we train a network, denoted as *MFIN-sup*, by applying the reconstruction loss function on the predicted motion fields with respect to *ground-truth* motion fields obtained from the gold-standard registration algorithm (Vishnevskiy et al., 2016) (Eq. (11)). Note that no additional regularization is applied here as the ground truth motion fields are already regularized. Such supervised training has been previously applied for optical flow estimation, like in (Fischer et al., 2015; Ilg et al., 2017; Ranjan and Black, 2017).

$$L_{\text{recon,MFIN-sup}} = L\Big(\mathbf{F}^{gt}_{t\to t-1}, \mathbf{F}_{t\to t-1}\Big) + L\Big(\mathbf{F}^{gt}_{t\to t+1}, \mathbf{F}_{t\to t+1}\Big) \tag{11}$$

3.5. 4D Reconstruction

In this section, we describe the 4D reconstruction procedure of Von Siebenthal et al. (2007). As stated before, the original acquisition scheme proceeds by acquiring alternately navigators and 2D data slices, with the navigators tagging their neighbouring data slices to the phase of the breathing cycle. Retrospectively, a MRI volume is reconstructed for each acquired data slice D_t^p, enclosed by navigators (N_t, N_{t+1}). This requires finding all slices in the data volume, each acquired at the same breathing phase as D_t^p. To this end, the data slice at location q $(\neq p)$ is computed as the average of five data slices D_s^q whose navigators (N_s, N_{s+1}) are most similar in mean liver position to (N_t, N_{t+1}). The mean liver positions are computed by registering each acquired navigator to a pre-determined reference navigator using (Hartkens et al., 2002), the same registration method as in Von Siebenthal et al. (2007). The 4D MRIs, thus reconstructed either using all true navigators or with alternate navigators obtained via interpolation, are compared for evaluating the interpolation methods.

4. Experiments and results

4.1. Dataset

We carry out our experiments on 2D navigator images from an abdominal 4D MRI dataset consisting of 14 subjects. The interleaved acquisition of navigator and data slices (Von Siebenthal et al., 2007) was done on a 1.5T Philips Achieva scanner using a 4-channel cardiac array coil, a balanced steady-state free precession sequence, SENSE factor 1.7, 70^o flip angle, 3.1 ms TR, and 1.5 ms TE. The images have a spatial resolution of $1.33 \times 1.33\,\text{mm}^2$, slice thickness of 5mm and temporal resolution of 2.4–3.1 Hz. For each subject, there are between 4000 and 6000 navigators, continuously acquired in several blocks of 7 to 9 min and with 5 min resting periods in between. Expert-annotated landmarks for two liver vessels per image were available for 10% randomly selected images for 7 out of the 14 subjects. We train on the remaining 7 subjects and use the 7 subjects with expert annotations as test subjects, so that the accuracy of the predicted motion fields could be evaluated as described in Section 4.3. The hyper-parameters for the experiments were either chosen as in Karani et al. (2017) or were determined empirically using a smaller experimental setup with training data as 4 blocks of one training subject and the test data as another block of the same subject. Thus, the test data was not involved at all during the training phase.

4.2. Implementation details

We implement MFIN-s, MFIN and MFIN-c (Fig. 3) as networks with an encoder-decoder like structure. All networks consist of an initial block of shared layers, followed by 1, 2 and 3 separate sub-networks for MFIN-s, MFIN and MFIN-c respectively. The shared layers include the entire encoder / contracting path as well as two upscaling layers from the decoder. Each sub-network consists of two convolutional layers, followed by a bilinear upsampling and then a final convolutional layer to obtain a motion field. There is no activation function at the end of the last convolutional layer to allow for both positive and negative flow values. All other convolutional layers are following by a ReLU activation function. Finally, a warping layer transforms the known image via the corresponding predicted motion field to provide the interpolated image. The structure of MFIN-sup is the same as that of MFIN, except that the loss function is applied on the motion fields predicted before the warping layer.

The filter sizes and number of feature maps are empirically set to $(\mathrm{f}_1, \mathrm{f}_2, \ldots, \mathrm{f}_8) = (7{,}5{,}3{,}3{,}3{,}3{,}3{,}3)$ and $(\mathrm{n}_2, \mathrm{n}_2, \ldots, \mathrm{n}_8) = (16{,}32{,}64{,}64{,}32{,}32{,}16)$. For interpolating a navigator at any time point, two known navigators from the past and the future are provided as inputs, i.e. $\mathrm{n}_1 = 4$. The output of each sub-network before the warping layer is a 2D flow vector for each pixel, i.e. $\mathrm{n}_9 = 2$. Following (Karani et al., 2017), we set the batch size to 64 and use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 1e-4. The only pre-processing step is block-wise linear normalization of the images to their 2 to 98%tile range. Following (Wang et al., 2004), the hyperparameters of the SSIM loss are set to $c_1 = 0.0001$, $c_2 = 0.0009$ and patch size to 11×11. The weights for the regularizers in the loss functions are empirically set as $\{\lambda_1 = 0.001, \lambda_2 = 0.0005\}$ and $\{\lambda_1 = 0.1, \lambda_2 = 0.05\}$ when using the L_2 loss and the $\mathrm{L}_{\mathrm{SSIM}}$ loss respectively for the reconstruction loss.

4.3. Evaluation metrics

The choice of appropriate evaluation metrics is crucial to correctly compare competing solutions. Here, we list several metrics and discuss their suitability for evaluating interpolation in the setting of navigator slices for 4D MR reconstruction.

M1 **Root-mean-squared-error (RMSE)**: RMSE and mean-absolute-error are generally used for evaluation in regression problems. Despite their wide-spread use, they also have some drawbacks for measuring image fidelity. Firstly, they are measured pixel-wise and hence, do not encode structural information. Secondly, these metrics are not invariant to intensity scaling. In the case of interpolation of navigators, we are only interested in the correct location of the structures in the image. Thus, a shift in the average intensity should not be punished. This might be relevant in cases when the navigator is affected by saturation, as in acquisitions with interleaved navigator and data slices. Finally, these metrics are sensitive to noise and noise is inherently present in MR images. The L_2 distance between a given ground truth image and a smoothened (denoised) interpolated image would be smaller than the case where the interpolated image had different noise values that the ground truth image.

M2 **Structural Similarity (SSIM) Index**: SSIM encodes local correlations between image patches along with their intensity similarity. Thus, it may be expected to be invariant to intensity scaling and more robust to image distortions than RMSE. On the other hand, it is not robust to noise (Ndajah et al., 2010) and also does not directly assess important clinically relevant aspects like correct organ position or accurate motion estimation.

M3 **Residual motion (ResMot)**: We register the interpolated image to the ground truth image via a gold standard (gs) image registration algorithm (linearly interpolated grid of control points, optimized for local correlation coefficient and total variation regularization) (Vishnevskiy et al., 2016), and compute an error motion field $\mathbf{F}^{\mathrm{gs}}_{t\to\hat{t}}$. The mean magnitude of this motion field could be relevant for 4D reconstruction as it measures the mismatch in organ positions.

M4 **Error in motion to a reference image (RefMotErrIm)**: For 4D reconstruction, each navigator image is registered to a reference image to estimate the position of the structure of interest in the navigator. To measure the error introduced in this step due to interpolation, we compute the difference of motion fields obtained via GS registration between the reference image and either (i) an interpolated navigator or (ii) the corresponding true navigator, thus obtaining flow fields $\mathbf{F}^{\mathrm{gs}}_{t\to\mathrm{ref}}$ or $\mathbf{F}^{\mathrm{gs}}_{\hat{t}\to\mathrm{ref}}$ respectively. From these flow fields, we compute two evaluation measures: the mean difference in their magnitudes over the entire image (RefMotErrIm) or only over the structure of interest, the liver in this application (RefMotErrImLiver). Note that the latter is the more relevant measure for reconstructing 4D MRIs to capture liver motion and can be determined as motion fields are defined with respect to the reference image, which has an associated liver segmentation.

M5 **Error in 4D reconstruction and extracted 3D Motion**: The final and the most application relevant evaluation criterion for the interpolation methods is the quality of the reconstructed 4D MRIs. As the reconstruction procedure (Section 3.5) is computationally very expensive, this is done only for the best-performing among the interpolation proposed methods (according to the other evaluation metrics) and compared to baselines of black-box interpolation (SCIN) and reconstruction using all true navigators. Further, we also compare the 3D liver motion extracted from the reconstructed 4D MRIs from the different methods. As in Von Siebenthal et al. (2007); Tanner et al. (2016a,b), the motion is extracted by an intensity-based image registration method based on maximizing normalized cross correlation within the reference liver region and a B-spline transformation model (Hartkens et al., 2002). The error of this registration method is reported as below one voxel (Von Siebenthal et al., 2007), based on visual inspection. The motion is sampled at corresponding locations from all test subjects using either a landmark-based approach (right liver lobe) or a shape-based approach (whole liver) (Tanner et al., 2016a).

The interpolation formulation in MFIN provides an unsupervised estimation of the motion fields between the interpolated image and its neighbours. We use the following measures for evaluating the accuracy for these motion fields.

M6 **Using the estimated motion fields for determining positions of interpolated images (RefMotErrFl)**: As mentioned before, the crucial step in 4D reconstruction is to estimate the position of the structure of interest in each navigator. This is usually done by registering each navigator to a reference image and is the most time consuming step in the reconstruction based on (Von Siebenthal et al., 2007). Since the interpolation of a navigator N_t provides the motion field $\mathbf{F}_{t\to t+1}$, we can use it to reduce the number of navigator registrations by almost half. This can be achieved by inverting the predicted motion field $\mathbf{F}_{t\to t+1}$ to get $\mathbf{F}_{t+1\to t}$ and then

Table 1
Quantitative results. SCIN: baseline (Karani et al., 2017), MFIN: Interpolation via bi-directional motion field prediction, MFIN-c: MFIN with cycle consistency constraint, MFIN-s: MFIN with only single direction motion field prediction (from t to $t-1$), MFIN-sup: MFIN trained in a supervised manner, with ground truth motion fields obtained using (Vishnevskiy et al., 2016). %ile refers to 5 percentile values for SSIM and 95 percentile otherwise. **Bold** font marks most important evaluation metrics for 4D reconstruction and results within 5% (for SSIM, 2.5%) of best values, which are $\underline{\text{underlined}}$. Asterisk (*) denotes that the landmark error for MFIN-s has been computed from two separate single-directional MFIN-s networks.

	SCIN-		MFIN-s-		MFIN-		MFIN-c-		MFIN-	
	L_2		L_2		L_2		L_2		sup-L_2	
Evaluation Metric	mean	%ile	mean	%ile	mean	%ile	mean	%ile	mean	%ile
RMSE	$\underline{\mathbf{8.78}}$	$\underline{\mathbf{11.83}}$	10.38	13.99	10.35	13.90	10.23	13.74	11.07	15.21
SSIM [%]	**82.08**	**77.15**	79.45	74.26	79.55	74.35	79.61	74.53	78.78	72.87
ResMot [mm]	$\underline{\mathbf{0.30}}$	$\underline{\mathbf{0.55}}$	0.37	0.70	0.37	0.67	0.36	0.67	0.48	0.90
RefMotErrIm	$\underline{\mathbf{0.53}}$	$\underline{\mathbf{0.98}}$	0.58	1.06	0.58	1.05	0.57	1.04	0.69	1.23
RefMotErrImLiver	0.70	1.43	0.72	1.47	0.72	1.44	0.70	**1.42**	0.86	1.73
MotErrFl [mm]	–	–	0.66	1.24	0.65	1.24	0.66	1.23	$\underline{\mathbf{0.59}}$	$\underline{\mathbf{1.06}}$
RefMotErrFl [mm]	–	–	**0.83**	1.65	**0.83**	1.64	**0.84**	1.64	$\underline{\mathbf{0.80}}$	$\underline{\mathbf{1.52}}$
RefMotErrFlLiver	–	–	0.83	1.66	0.83	1.63	0.83	1.66	0.92	1.80
LandmarkErr [mm]	–	–	***0.96**	*1.98	0.98	**1.88**	**0.93**	1.97	1.13	2.41

	SCIN-		MFIN-s-		MFIN-		MFIN-c-	
	L_{SSIM}		L_{SSIM}		L_{SSIM}		L_{SSIM}	
Evaluation Metric	mean	%ile	mean	%ile	mean	%ile	mean	%ile
RMSE	**9.05**	**12.03**	10.33	14.32	10.30	14.09	10.28	14.06
SSIM [%]	$\underline{\mathbf{82.72}}$	$\underline{\mathbf{78.01}}$	79.65	74.30	79.81	74.67	79.87	74.78
ResMot [mm]	$\underline{\mathbf{0.30}}$	**0.58**	0.34	0.68	0.37	0.68	0.37	0.67
RefMotErrIm	**0.54**	**1.00**	0.59	1.08	0.59	1.08	0.58	1.05
RefMotErrImLiver	$\underline{\mathbf{0.66}}$	**1.37**	0.73	1.43	0.70	**1.39**	**0.68**	$\underline{\mathbf{1.35}}$
MotErrFl [mm]	–	–	0.68	1.27	0.66	1.25	0.64	1.23
RefMotErrFl [mm]	–	–	0.85	1.67	**0.84**	1.66	**0.82**	1.63
RefMotErrFlLiver	–	–	0.84	1.62	**0.80**	**1.57**	$\underline{\mathbf{0.78}}$	$\underline{\mathbf{1.53}}$
LandmarkErr [mm]	–	–	***0.94**	*$\underline{\mathbf{1.79}}$	**0.93**	**1.85**	$\underline{\mathbf{0.92}}$	**1.81**

composing it with $\mathbf{F}^{gs}_{ref\rightarrow t+1}$ (obtained by registering N_{t+1} to the reference image) to estimate $\mathbf{F}^{gs}_{ref\rightarrow t}$. The error in the estimation ($||\mathbf{F}^{gs}_{ref\rightarrow t}$ - $\mathbf{F}_{t+1\rightarrow t} \odot \mathbf{F}^{gs}_{ref\rightarrow t+1}||_2$) serves as a measure of the accuracy of the predicted motion field $\mathbf{F}_{t\rightarrow t+1}$. As with measure M4, M6 can also be computed either over the entire image (RefMotErrFl) or only over the structure on interest (RefMotErrFlLiver).

M7 **Error in predicted motion field (MotErrFl)**: The evaluation measures from M6 include not only the error because of the wrong prediction of $\mathbf{F}_{t+1\rightarrow t}$ but also errors due to inconsistencies of the involved GS registrations i.e. if $||\mathbf{F}^{gs}_{ref\rightarrow t}$ - $\mathbf{F}^{gs}_{t+1\rightarrow t} \odot \mathbf{F}^{gs}_{ref\rightarrow t+1}||_2 > 0$. To discern these two effects, we directly calculated the error of the predicted flow field by comparing $\mathbf{F}^{gs}_{t+1\rightarrow t}$ and $\mathbf{F}_{t+1\rightarrow t}$.

M8 **Landmark error (LandmarkErr)**: Another method to evaluate the accuracy of the motion fields is to compute the landmark errors for the cases where we have expert annotations on consecutive navigators.

4.4. Experiments

An overview of our experiments is as follows. Section 4.4.1 describes the comparison between the performance of interpolation via motion field prediction against the black-box interpolation network (Karani et al., 2017). Next, in Section 4.4.2, we discuss the effect of the cycle consistency constraint. The analysis as to whether the bi-directional field prediction indeed provides a benefit over predicting the motion field only in a single direction is presented in Section 4.4.3. Section 4.4.4 describes the evaluation of the supervised training with ground truth motion fields. Finally, the effect of the interpolated navigators on the slice sorting leading to the 4D MRI reconstruction is analyzed in Section 4.4.5.

The interpolation networks SCIN, MFIN, MFIN-c and MFIN-s were each trained separately with two different functions forthe reconstruction loss: L_2 and L_{SSIM}. We denote these networks as SCIN-L_2, MFIN-c-L_{SSIM}, etc. For MFIN-sup, we employ the L_2 loss between the predicted and ground truth motion fields. Table 1 summarizes quantitative results of the navigator interpolations in terms of the evaluation metrics discussed in Section 4.3.

4.4.1. Interpolation via motion field prediction

We first compare the performance of the MFIN network with the baseline black-box interpolation network, SCIN. The following observations can be made about the quantitative results about this comparison:

- In terms of the intensity-based metrics (RMSE and SSIM) as well as the registration-based metrics (ResMot and RefMotErrImLiver), SCIN performs better than MFIN, for both training loss functions. However, as discussed in Section 4.3, the intensity-based metrics might not be appropriate for measuring interpolation performance. As well, ResMot might be artificially reduced for SCIN, as its blurring and denoising property is likely to reduce gradients of the image similarity measure optimized during registration. RefMotErrImLiver is the most relevant evaluation measure for the application of reconstruction of 4D MRIs. Even with respect to this measure, SCIN performs better than MFIN. Thus, even though the images interpolated by MFIN appear sharper, the quantitative results indicate that the accuracy of the predicted motion fields needs to be further improved.
- For both SCIN and MFIN, training with L_{SSIM} leads to lower errors with respect to RefMotErrImLiver than training with the L_2 loss.
- The flow-based error metrics, MotErrFl, RefMorErrFl and the error on manually annotated landmarks cannot be computed for

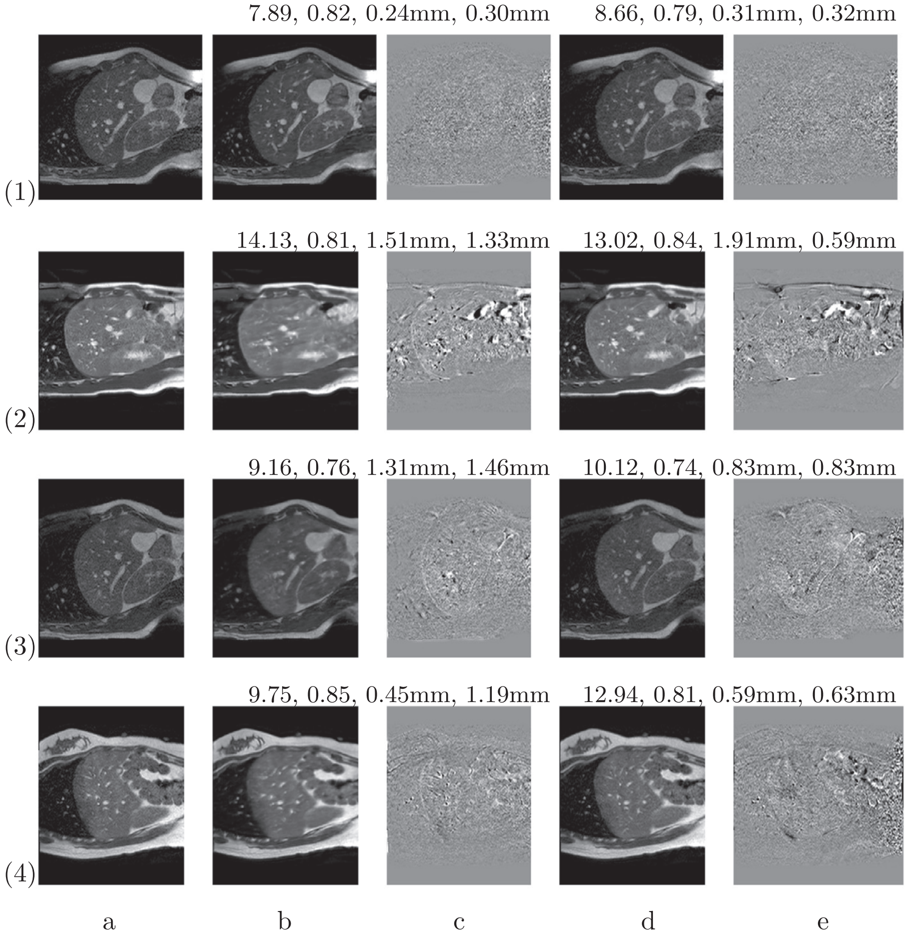
7.89, 0.82, 0.24mm, 0.30mm
8.66, 0.79, 0.31mm, 0.32mm
(1)
14.13, 0.81, 1.51mm, 1.33mm
13.02, 0.84, 1.91mm, 0.59mm
(2)
9.16, 0.76, 1.31mm, 1.46mm
10.12, 0.74, 0.83mm, 0.83mm
(3)
9.75, 0.85, 0.45mm, 1.19mm
12.94, 0.81, 0.59mm, 0.63mm
(4)
a
b
c
d
e

Fig. 4. Columns: (a) ground truth images, (b) SCIN-L$_{\text{SSIM}}$ results, (c) difference (b-a), (d) MFIN-c-L$_{\text{SSIM}}$ results, (e) difference (d-a). Rows: (1) low motion case, (2)-(4) high motion cases, where MFIN-c produces visually (2) much better, (3) slightly better and (4) worse structure alignment than SCIN. RMSE, SSIM, RefMotErrIm, RefMotErrImLiver values are stated over the respective image pairs.

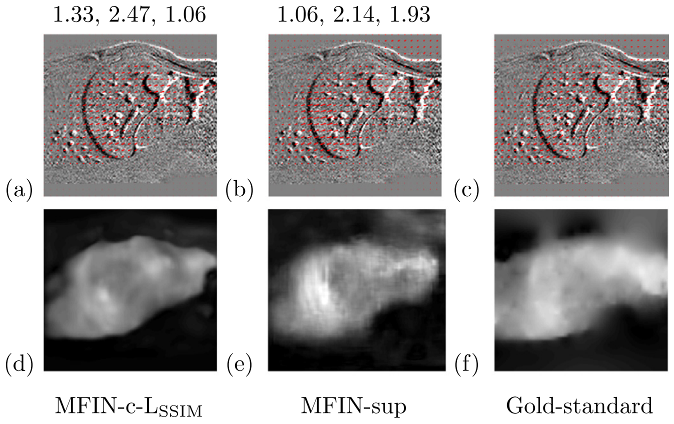

Fig. 5. (a-c) Motion field $\mathbf{F}_{t\to t+1}$ overlaid on $N_{t+1} - N_t$ from (a) MFIN-c-L$_{\text{SSIM}}$, (b) MFIN-sup, and (c) gold standard registration ($\mathbf{F}^{gs}_{t\to t+1}$). (d-f) Motion magnitude images corresponding to top row. MotErrFl, RefMotErrFl, RefMotErrFlLiver values [in mm] are stated over the respective images.

SCIN as it does not provide an estimation of the motion field between the interpolated image and its neighbours.

4.4.2. Effect of cycle consistency constraint

After incorporating the cycle consistency constraint, the difference in mean performance between SCIN-L$_{\text{SSIM}}$ and MFIN-c-L$_{\text{SSIM}}$ with regard to RefMotErrImLiver is 0.02mm. To test whether this difference in performance affects the reconstruction, we computed the error incurred in the data slice sorting that follows the navigator position determination step in 4D reconstruction. In the case where all ground truth navigators are used, the discrepancy between a given navigator and the closest navigator corresponding to another data slice is, on average, 1.48mm. This is much larger than the difference in RefMotErrImLiver between SCIN and MFIN-c or even MFIN. We thus infer that the increase in RefMotErrImLiver for MFIN-c or even for MFIN as compared to SCIN may not affect the reconstruction. Note that RefMotErrImLiver is higher than RefMotErrIm because the average motion magnitude to the reference is higher in the liver (mean 5.13, 95% 11.81 mm) than for the average for the whole image (mean 3.40, 95% 8.21 mm).

4.4.3. Effect of Bi-directional motion field prediction

It has been suggested in the literature that multi-task learning potentially creates inductive biases that may be beneficial for the individual tasks (Caruana, 1998). To investigate whether this hypothesis holds in our application as well, we trained an ablated version of MFIN with only one sub-network to predict the motion field in only the backward direction (from t to $t-1$). All error metrics for MFIN-s in Table 1, except M8, were computed with respect to this network. In order to compute M8 (landmark error) with respect to all consecutive slices with manual annotations, we additionally trained another network with single-directional motion field prediction in the forward direction (from t to $t+1$). We observed no substantial difference between single-direction and bi-directional motion field predictions. Nonetheless, bi-directional

Table 2
Effect of interpolated navigators by registration, SCIN-L_2 or MFIN-c-L_{SSIM} (all $T = 2$) on 4D MRI reconstruction (RMSE) and 3D motion extraction. Results within 5% to underlined lowest error are marked in **bold** font.

	RMSE		3D Motion Error [mm]			
	(image)		(whole liver)		(right liver lobe)	
Method	mean	95%	mean	95%	mean	95%
Registration	**3.87**	7.48	**0.79**	**2.03**	**0.64**	**1.57**
SCIN-L_2	4.17	**6.89**	**0.79**	**2.03**	**0.65**	**1.56**
MFIN-c-SSIM	**3.75**	**7.19**	**0.78**	**2.00**	**0.64**	**1.54**

motion field predictions are required when enforcing the cycle consistency constraint.

4.4.4. Benefit of supervised training with ground truth motion fields

The network trained to predict the gold-standard motion fields, MFIN-sup, provides the best performance when evaluating the whole image flow fields (MotErrFl, RefMotErrFl) as it is explicitly trained for this. However, this does not seem to be beneficial for keeping errors within the liver low (RefMotErrFlLiver, landmark error) or for providing well interpolated navigators. This method may be potentially affected by the errors in the gold standard motion fields such as non-zero motion in the background (see Fig. 5).

4.4.5. Effect of interpolated navigators on 4D reconstruction

We reconstruct 4D MRIs for four scenarios: (i) using all the true navigators, (ii-iv) using alternate true and interpolated navigators obtained by (ii) a registration-based interpolation method, used as a baseline in (Karani et al., 2017), (iii) SCIN-L_2 and (iv) MFIN-c-L_{SSIM}. Table 2 lists the 4D reconstruction performance of the three interpolation methods with respect to the 4D MRI generated using all true navigators. Most different 4D MRIs are reconstructed, on average, by SCIN-L_2. However, very similar 3D motion error values are achieved with all three methods. These results indicate that the various evaluation metrics discussed before may be sensitive to differences between interpolated images that do not highly influence the 4D reconstruction.

As discussed in Section 4.3 [M6], the motion fields provided by the interpolation network may be used to halve the number of navigator registration required for the 4D reconstruction. This can lead to substantial time savings as several hundred navigator registrations are involved in the reconstruction and each 2D registration requires 2.08 seconds, while the transformation inversion discussed in Section 4.3 [M6] needs only 0.65 seconds per image.

4.4.6. Qualitative results

We observed no large qualitative differences in the performances of MFIN and MFIN-c for either of the two loss functions. Since, MFIN-c-L_{SSIM} provides the best quantitative results, we show interpolated images from this method and compare them against SCIN-L_{SSIM} in Fig. 4. Both methods perform well when the motion between the neighbouring images is low. This is reflected in the absence of any structures in the error images in Fig. 4-(1). However, RMSE is lower for SCIN because it produces a denoised interpolated image, while MFIN carries over the noise pattern from the neighbouring known image. Whenever there exists high motion between the images being interpolated, SCIN produces blurry images and often distorts image structures. This can be observed in cases (2)-(4) in Fig. 4. For all these cases, MFIN-c (and also MFIN) produces sharp images and largely preserves structures in the images. Fig. 4-(2) shows a case where MFIN-c additionally has a much better performance with respect to image alignment. Fig. 4-(3) shows a representative case, with small improvement in image alignment, yet worse RMSE and SSIM values for MFIN-c. Finally, Fig. 4-(4) shows a case, where MFIN-c produces worse alignment of structures than SCIN near the gall-bladder (top-right), but better alignment in order regions such as the blood vessels near the liver edge. While it is hard to observe a clear relationship between the registration-based error metrics and the qualitative results, registration error over the liver (RefMotErrImLiver) seems to be reduced for cases with lower organ misalignment (e.g. Fig. 4-(2,3)).

Fig. 5 shows a comparison of a representative motion field predicted by MFIN-c-L_{SSIM} and MFIN-sup with that computed via the GS registration algorithm. We can see that the motion field produced by MFIN-c is smooth and has sharper motion boundaries. The reason for this might be that the used registration is more regularized due to its parametric model, where motion is defined by a grid of control points with 4x4 pixel spacing and linearly interpolated in between. This might also explain the higher error in evaluation of the flow field predicted by the network over the whole image (RefMotErrFl) than only over the liver (RefMotErrFlLiver). Fig. 5 also shows that MFIN-sup learns to provide a flow field which has a pattern more similar to the gold standard field, with motion in the background and less sharp boundaries.

5. Conclusion

In this article, we proposed a framework for temporal image interpolation that incorporates the prior knowledge that changes in the images over time are caused by the motion of the visible structures in the images. We showed that this approach preserves structures in the images and produces sharper interpolated images than direct interpolation in the intensity space. Such behaviour would likely be advantageous for potential downstream tasks on temporal data, for example, tracking of segmentation labels. We also showed the advantage of SSIM as a loss measure for optimization as compared to the L_2 loss and introduced a cyclic consistency constraint between the bi-directional motion fields to further improve interpolation performance. Finally, we evaluated the interpolation performance extensively and provided a detailed analysis about the suitability of several evaluation metrics. Our experiments highlight the importance of employing application-specific metrics for proper evaluation.

Another important benefit of the proposed method is that it provides a free estimation of the motion fields between the known and interpolated images. In the particular case of navigated 2D multi-slice imaging, these motion fields could be possibly utilized to lower the number of registration required for the 4D reconstruction, potentially leading to large time savings. Such time savings could be further exaggerated if the interpolation method is extended to 3D interpolations. Also, although we presented results in the setting of 4D MRI reconstruction, the method may be extended to other scenarios where the content of temporal sequences does not change over time.

References

Arnold, P., Preiswerk, F., Fasel, B., Salomir, R., Scheffler, K., Cattin, P., 2011. 3D organ motion prediction for MR-guided high intensity focused ultrasound. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, p. 623.

Baumgartner, C., Kolbitsch, C., McClelland, J., Rueckert, D., King, A., 2013. Groupwise simultaneous manifold alignment for high-resolution dynamic MR imaging of respiratory motion. In: Information Processing in Medical Imaging, p. 232.

Bert, C., Durante, M., 2011. Motion in radiotherapy: particle therapy. Phys. Med. Biol. 56(16), R113.

Cai, J., Chang, Z., Wang, Z., Paul Segars, W., Yin, F.-F., 2011. Four-dimensional magnetic resonance imaging (4d-mri) using image-based respiratory surrogate: a feasibility study. Med. Phys. 38 (12), 6384–6394.

Caruana, R., 1998. Multitask Learning. In: Learning to learn. Springer, p. 95.

Christensen, G.E., Johnson, H.J., 2001. Consistent image registration. IEEE Trans. Med. Imag. 20 (7), 568–582.

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: IEEE International Conference on Computer Vision, pp. 2758–2766.

Fischer, P., Dosovitskiy, A., Ilg, E., Häusser, P., Hazırbaş, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T., 2015. Flownet: learning optical flow with convolutional networks. arXiv:1504.06852.

Gass, T., Székely, G., Goksel, O., 2015. Consistency-based rectification of nonrigid registrations. J. Med. Imaging 2(1), 014005.

Goroshin, R., Mathieu, M., LeCun, Y., 2015. Learning to linearize under uncertainty. In: Advances in Neural Information Processing Systems, pp. 1234–1242.

Hartkens, T., Rueckert, D., Schnabel, J., Hawkes, D., Hill, D., 2002. VTK CISG registration toolkit: an open source software package for affine and non-rigid registration of single- and multimodal 3D images. In: Bildverarbeitung für die Medizin, p. 409.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T., 2017. Flownet 2.0: evolution of optical flow estimation with deep networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2.

Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks. In: Advances in Neural Information Processing Systems, pp. 2017–2025.

Jason, J., Harley, A., Derpanis, K., 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.

Jiang, H., Sun, D., Jampani, V., Yang, M., Learned-Miller, E., Kautz, J., 2017. Super slomo: high quality estimation of multiple intermediate frames for video interpolation. arXiv:1712.00080.

Karani, N., Tanner, C., Kozerke, S., Konukoglu, E., 2017. Temporal interpolation of abdominal MRIs acquired during free-breathing. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, p. 359.

Keall, P., Starkschall, G., Shukla, H., Forster, K., Ortiz, V., Stevens, C., Vedam, S., George, R., Guerrero, T., Mohan, R., 2004. Acquiring 4d thoracic ct scans using a multislice helical method. Phys. Med. Biol. 49 (10), 2053.

Kingma, D., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv:1412.6980.

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

Lambert, J., Suchowerska, N., McKenzie, D., Jackson, M., 2005. Intrafractional motion during proton beam scanning. Phys. Med. Biol. 50 (20), 4853.

Lee, G.-I., Park, R.-H., Song, Y.-S., Kim, C.-A., Hwang, J.-S., 2003. Real-time 3D ultrasound fetal image enhancment techniques using motion-compensated frame rate up-conversion. In: Medical Imaging, p. 375.

Li, G., Citrin, D., Camphausen, K., Mueller, B., Burman, C., Mychalczak, B., Miller, R.W., Song, Y., 2008. Advances in 4d medical imaging and 4d radiation therapy. Technol. Cancer Res. Treatment 7 (1), 67–81.

Liu, Z., Yeh, R.A., Tang, X., Liu, Y., Agarwala, A., 2017. Video frame synthesis using deep voxel flow. In: ICCV, pp. 4473–4481.

Long, G., Kneip, L., Alvarez, J., Li, H., Zhang, X., Yu, Q., 2016. Learning image matching by simply watching video. In: ECCV. Springer, p. 434.

Low, D.A., Nystrom, M., Kalinin, E., Parikh, P., Dempsey, J.F., Bradley, J.D., Mutic, S., Wahab, S.H., Islam, T., Christensen, G., et al., 2003. A method for the reconstruction of four-dimensional synchronized ct scans acquired during free breathing. Med. Phys. 30 (6), 1254–1263.

Mathieu, M., Couprie, C., LeCun, Y., 2015. Deep multi-scale video prediction beyond mean square error. arXiv:1511.05440.

Nam, T.-J., Park, R.-H., Yun, J.-H., 2006. Optical flow based frame interpolation of ultrasound images. In: Int Conf Image Analysis and Recognition, p. 792.

Ndajah, P., Kikuchi, H., Yukawa, M., Watanabe, H., Muramatsu, S., 2010. SSIM image quality metric for denoised images. In: Proc. 3rd WSEAS Int. Conf. on Visualization, Imaging and Simulation, p. 53.

Niklaus, S., Mai, L., Liu, F., 2017. Video frame interpolation via adaptive separable convolution. arXiv:1708.01692.

Patraucean, V., Handa, A., Cipolla, R., 2015. Spatio-temporal video autoencoder with differentiable memory. arXiv:1511.06309.

Ranjan, A., Black, M., 2017. Optical flow estimation using a spatial pyramid network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2.

Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv:1312.6034.

Snell, J., Ridgeway, K., Liao, R., Roads, B., Mozer, M., Zemel, R.S., 2017. Learning to generate images with perceptual similarity metrics. In: Image Processing (ICIP), 2017 IEEE International Conference on. IEEE, pp. 4277–4281.

Srivastava, N., Mansimov, E., Salakhutdinov, R., 2015. Unsupervised learning of video representations using LSTMs. In: ICML, p. 843.

Tanner, C., Yang, M., Samei, G., Székely, G., 2016. Influence of inter-subject correspondences on liver motion predictions from population models. In: IEEE Int. Symposium on Biomedical Imaging, pp. 286–289.

Tanner, C., Zur, Y., French, K., Samei, G., Strehlow, J., Sat, G., McLeod, H., Houston, G., Kozerke, S., Székely, G., Melzer, A., 2016. In vivo validation of spatio-temporal liver motion prediction from motion tracked on MR thermometry images. Int. J. Comput. Assist. Radiol. Surg. 11 (6), 1143–1152.

Tryggestad, E., Flammang, A., Han-Oh, S., Hales, R., Herman, J., McNutt, T., Roland, T., Shea, S., Wong, J., 2013. Respiration-based sorting of dynamic MRI to derive representative 4D-MRI for radiotherapy planning. Med. Phys. 40(5), 051909.

Vishnevskiy, V., Gass, T., Szekely, G., Tanner, C., Goksel, O., 2016. Isotropic total variation regularization of displacements in parametric image registration. IEEE T. Med. Imag. 36(2), 385.

Von Siebenthal, M., Székely, G., Gamper, U., Boesiger, P., Lomax, A., Cattin, P., 2007. 4D MR Imaging of respiratory organ motion and its variability. Phys. Med. Biol. 52, 1547.

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.

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

Yang, X., Kwitt, R., Styner, M., Niethammer, M., 2017. Quicksilver: fast predictive image registration–a deep learning approach. Neuroimage 158, 378–396.

Yeh, R., Liu, Z., Goldman, D., Agarwala, A., 2016. Semantic facial expression editing using autoencoded flow. arXiv:1611.09961.

Zhang, L., Karani, N., Tanner, C., Konukoglu, E., 2018. Temporal interpolation via motion field prediction. arXiv:1804.04440.

Zhang, W., Brady, J., Becher, H., Noble, J., 2011. Spatio-temporal (2D+ t) non-rigid registration of real-time 3D echocardiography and cardiovascular MR image sequences. Phys. Med. Biol. 56(5), 1341.

Zhou, T., Tulsiani, S., Sun, W., Malik, J., Efros, A.A., 2016. View synthesis by appearance flow. In: European Conference on Computer Vision. Springer, pp. 286–301.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv:1703.10593.