Population-based 3D respiratory motion modelling from convolutional autoencoders for 2D ultrasound-guided radiotherapy

Article *in* Medical Image Analysis · October 2021

DOI: 10.1016/j.media.2021.102260

4 authors, including:

Liset Vázquez Romaguera
Polytechnique Montréal
29 PUBLICATIONS **96** CITATIONS

SEE PROFILE

William Trung Le
University of Montreal Hospital Research Centre
13 PUBLICATIONS **38** CITATIONS

SEE PROFILE

Samuel Kadoury
Polytechnique Montréal
204 PUBLICATIONS **4,586** CITATIONS

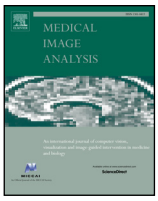
SEE PROFILE

Medical Image Analysis 75 (2022) 102260

Contents lists available at ScienceDirect

Medical Image Analysis

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



Population-based 3D respiratory motion modelling from convolutional autoencoders for 2D ultrasound-guided radiotherapy

Tal Mezheritsky[a,*], Liset Vázquez Romaguera[a], William Le[b], Samuel Kadoury[a,b]

[a] MedICAL Laboratory, École Polytechnique de Montréal, Montréal, Canada
[b] CHUM Research Center, Montréal, Canada

ARTICLE INFO

Article history:
Received 19 March 2021
Revised 29 September 2021
Accepted 1 October 2021
Available online 9 October 2021

MSC:
41A05
41A10
65D05
65D17

Keywords:
Motion modelling
Ultrasound-guided radiotherapy
Deformable registration
Liver cancer
Convolutional autoencoders

ABSTRACT

Radiotherapy is a widely used treatment modality for various types of cancers. A challenge for precise delivery of radiation to the treatment site is the management of internal motion caused by the patient's breathing, especially around abdominal organs such as the liver. Current image-guided radiation therapy (IGRT) solutions rely on ionising imaging modalities such as X-ray or CBCT, which do not allow real-time target tracking. Ultrasound imaging (US) on the other hand is relatively inexpensive, portable and non-ionising. Although 2D US can be acquired at a sufficient temporal frequency, it doesn't allow for target tracking in multiple planes, while 3D US acquisitions are not adapted for real-time. In this work, a novel deep learning-based motion modelling framework is presented for ultrasound IGRT. Our solution includes an image similarity-based rigid alignment module combined with a deep deformable motion model. Leveraging the representational capabilities of convolutional autoencoders, our deformable motion model associates complex 3D deformations with 2D surrogate US images through a common learned low dimensional representation. The model is trained on a variety of deformations and anatomies which enables it to generate the 3D motion experienced by the liver of a previously unseen subject. During inference, our framework only requires two pre-treatment 3D volumes of the liver at extreme breathing phases and a live 2D surrogate image representing the current state of the organ. In this study, the presented model is evaluated on a 3D+t US data set of 20 volunteers based on image similarity as well as anatomical target tracking performance. We report results that surpass comparable methodologies in both metric categories with a mean tracking error of 3.5±2.4 mm, demonstrating the potential of this technique for IGRT.

1. Introduction

Radiation therapy (RT) is used in more than 50% of cancer patients to treat and control disease progression (Jaffray, 2015). External beam radiotherapy (EBRT), a specific modality of RT, uses an external radiation source and collimators to deliver precise doses of radiation to the tumor site from different orientations around the patient's body. The goal of EBRT is to deliver enough radiation to damage the genetic material of cancerous cells, thus disabling them from dividing and growing the cancerous tumor further (Baumann et al., 2008). However, radiation is not only harmful to cancerous cells, it can also damage healthy cells (Dormand et al., 2005), making the precision of RT delivery systems crucial, especially for organs at risk. In the case of EBRT, the most complex sites to treat are the ones that experience severe motion induced by the patient's breathing. Indeed, respiratory motion poses great challenges to the administration during EBRT due to large motion organs such as the liver (Hawkes et al., 2005). Generally, the motion of abdominal organs due to respiration is most prominent in the superior-inferior (SI) direction with no more than 2 mm of motion in the anterior-posterior (AP) and lateral directions (Keall et al., 2006). The average displacement in the SI direction of the liver during shallow breathing varies between 13–25 mm (Davies et al., 1994; Weiss et al., 1972). The average displacement for deep breathing was determined to be 55 mm by Suramo et al. (1984) This forces radio-oncologists to increase the treatment margins to reduce the probability of cancer recurrence, thus increasing toxicity to healthy tissues (Keall et al., 2006). In an attempt to minimize the negative effects of respiratory motion on the efficiency of EBRT, a variety of respiratory motion management techniques have been proposed and used in clinical settings.

For respiratory gating approaches, the treatment is administered only within a predefined range of the patient's respiratory

* Corresponding author.
E-mail address: tal.mezheritsky@polymtl.ca (T. Mezheritsky).

https://doi.org/10.1016/j.media.2021.102260
1361-8415/

cycle using breath-holds. On the other hand, the forced shallow breathing technique does not require the patient to temporarily stop breathing, however it reduces the amplitude of respiratory motion by applying pressure to the patient's abdomen (Keall et al., 2006). While improving the precision of EBRT, the aforementioned techniques bear limitations such as increased treatment time and physical discomfort to the patient. For image-guided radiotherapy (IGRT), the aim is to use imaging to track the treatment target at all times during the administration of radiation to the tumor site. As the target is tracked, the delivery system adjusts its beam to account for the displacement of the tumor inside the patient's body. Therefore, IGRT has the potential of reducing the amount of damage caused to healthy tissues due to large treatment margins, all while allowing the patient to breath freely during the procedure (Brock and Dawson, 2010).

A wide range of imaging modalities can be used in the context of IGRT. X-ray imaging with or without the implantation of fiducial markers is often used in clinical practice. An early work by Schweikard et al. (2000) presented a method to compensate for respiratory motion during radiotherapy by using a correlation model between the displacement of implanted X-ray markers and infrared skin surface images. The main drawback of using X-ray imaging for motion compensation is that the additional radiation dose X-ray imparts reduces the imaging frame-rate that can be used. Similarly, cone-beam computed tomography (CBCT) cannot be used in real time during treatment due to significant exposure to ionizing radiation. In recent years systems that use MRI for IGRT have emerged, however they aren't widely available yet (Western et al., 2015). In contrast, ultrasound (US) is a non-ionizing, portable and inexpensive medical imaging modality that circumvents most of the disadvantages of other imaging modalities within the scope of IGRT. As current US system are capable of 2D, 3D and 3D+t imaging, they can be used both in the planning and treatment stages of the RT workflow (Fontanarosa et al., 2015). For example, Sawada et al. (2004) presented a novel respiratory gated radiation therapy system which allowed to trigger the treatment beam using image correlation between US images acquired on 3 orthogonal planes and a reference volume. Although the study was only carried out on a phantom setup, it presented US imaging as a viable modality for tracking moving targets during IGRT treatments.

US imaging can also be used to track the motion of the prostate using 3D to 2D US registration. Gillies et al. (2017) proposed a real-time automatic motion correction algorithm for fusion-based prostate biopsy systems. Their approach was able to achieve an average error of 1.6 $\pm$ 0.6 mm. Selmi et al. (2018) proposed a modified version of the iterative closest point algorithm (ICP) to perform real-time navigation in computer-assisted prostate biopsy systems. Matched features extracted from live 2D US images and a 3D US volume are used to drive the optimization process. The authors reported and average target registration error of 3.91 $\pm$ 3.22 mm.

Samei et al. (2018) proposed a real-time deformable registration technique. The proposed gradient descent technique is applied between a thin volume consisting of consecutive intraoperative 2-D transrectal ultrasound (TRUS) images and a preoperative 3D TRUS volume. The reported average accuracy for a dataset of 11 patients was 0.72 mm with an initial target displacement of 4.62 mm.

Current US-based abdominal IGRT systems rely on 2D imaging to track targets during imaging, even though targets are known to experience complex 3D trajectories especially in organs such as the liver and lungs (Keall et al., 2006). Therefore, 3D US imaging can be useful in IGRT applications. Clinically available 3D US matrix-array probes provide complete anatomical information of the tissues surrounding the tumor target, still the acquisition frame-rate is significantly lower than in 2D US imaging and the considerable storage size of 3D volumes significantly increases processing and computing times, making it difficult to use for real-time IGRT applications (Western et al., 2015). As such, we present a hybrid solution employing both 2D and 3D US for US-guided EBRT, by learning the relationship between 2D images and 3D deformation fields for real-time inference of volumetric US imaging.

1.1. Related works

The task of tracking anatomical targets in 3D US has generated significant interest, leading to open challenges like CLUST15 (Luca et al., 2015), providing a common datasets to compare solutions both on 2D and 3D temporal US sequences. Shepard et al. (2017) proposed a block matching multi-step tracking approach where each step accounted for an increasingly finer level of motion. Ozkan et al. (2017) proposed a tracking technique based on supporter features surrounding the tracking target. By tracking the supporters, the tracking accuracy of the desired target was improved. Both approaches achieved sub-millimeter performance, however they were only tested on 2D images. Methods tested on 3D US data included Banerjee et al. (2015), registering a global point set across temporal volumes using block matching, followed with a 3D registration of a local point set around the anatomical landmark, while Royer et al. (2017) represented the 3D target as a model of tetrahedral cells and vertices. The internal and external motion of the target mesh were estimated using a mechanical model and an intensity based approach respectively. In general, local tracking methods share a common disadvantage in failing to provide information about the motion of surrounding tissues which could be useful for dose re-planning (McClelland et al., 2013).

Global tracking solutions, on the other hand, attempt to determine the new position of a target by providing the motion experienced by its surroundings and the treated organ as a whole. The expected output becomes a motion field that spans the entire volume, which can be used not only to track treatment targets but also to adjust treatment planning and dose calculation. Obtaining complex 3D motion fields by leveraging inputs of a lower dimension has been commonly achieved in the context of IGRT through the use of respiratory motion modelling (McClelland, 2013). Surrogate signals can be obtained through 1D signals such as spirometry, skin surface motion tracking or 2D images of the treated organ, which can be used during treatment to infer the 3D motion field experienced at the time of the procedure (McClelland, 2013). However, very few works focused on respiratory motion modelling for 3D US due to the inherent difficulties such as low image quality and presence of unique artifacts (McClelland et al., 2013). Nevertheless, respiratory motion modelling remains flexible in terms of modality choice, even allowing to use one modality as a surrogate for another in certain applications (Preiswerk et al., 2014).

A first group of global tracking solutions based on respiratory motion modelling are patient-specific, where before treatment, the acquisition of 3D+t data along with surrogate signals is performed on the patient. The 3D motion is then obtained through registration of the 3D+t data to a reference volume chosen at a certain respiratory phase. Several approaches establishing a correspondence between surrogate signals and motion fields have been proposed. Arnold et al. (2011) created an atlas of motion from 3D+t MRI data, which is recovered using a respiratory signal acquired during treatment. Noorda et al. (2016) acquired cine MRI slices at 6 positions across the liver and registered them to a reference 3D MRI volume to obtain a lookup table of extrapolated 3D deformation fields corresponding to a variety of liver states. Among the works on patient-specific respiratory motion models, principal component analysis (PCA) stands out as a reference, using a linear decomposition of the patient-specific 3D motion fields, which is recov-

ered using a surrogate signal, such as a 2D navigator (King et al., 2012). Other means to obtain PCA combination coefficients include maximizing image similarity between acquired surrogate and deformed reference volume slice (Harris et al., 2016; Stemkens et al., 2016; Pham et al., 2019) or sparse block matching (Ha et al., 2019). McClelland et al. (2017) proposed to unify the steps of motion calculation and establish the surrogate correspondences which are usually separated. Their general framework showed many advantages, however high computation time limits its use in real-time applications. The main drawback with these approaches is that patient-specific 3D+t data is needed in order to model the respiratory motion patterns, which is far from being widely accessible in all institutions.

The second group of global motion models, referred to as population-based or cross-population models, aims to capture a wider variety of motion fields by capturing motion variability across a population of patients. Samei et al. (2012) introduced the concept of exemplar models, where each patient in the data set was used to fit exemplar patient-specific models. For new patients, the obtained surrogate is compared to the exemplar models and an optimized linear combination of all the patient-specific models is obtained. Paganelli et al. (2018) proposed a global respiratory motion model that directly infers the complete 3D deformation field by extrapolating the registration of interleaved 2D MRI surrogates with the planning MRI volume. Just as subject-specific models, PCA is also widely used when constructing population based respiratory motion models (Boye et al., 2013; Tanner et al., 2016; Jud et al., 2017). Preiswerk et al. (2014) proposed to combine information from 2D US images with a PCA respiratory motion model to predict the 3D motion of the liver acquired using MRI. However, the main drawback of using PCA is the requirement of establishing inter-patient correspondences, which is a time-consuming and often inaccurate process.

Deep learning has allowed to explore new solutions for the problem of target tracking in IGRT. Local tracking solutions in 2D (Huang et al., 2019; Liu et al., 2020) and 3D (He et al., 2019) using convolutional neural networks (CNN) have achieved accuracies within 0.69-1.89 mm. Giger et al. (2018) proposed a subject-specific respiratory motion model based on conditional generative adversarial networks (cGAN). The cGAN learned to predict 3D deformations of MRI based on a simultaneously acquired 2D US surrogate. However this method was only validated on three subjects. Romaguera et al. (2020) introduced a global respiratory motion model based on CNN and convolutional long short-term memory (CLSTM) units to perform in-plane target tracking with up to 5 timesteps prediction. The model was validated on 3 imaging modalities (MRI, CT and US), however it can only be applied to 2D images. Mezheritsky et al. (2020) proposed a respiratory motion model to generate up-to-date US volumes by combining image features from a reference 3D volume and a current 2D US image. While the model showed promise, its validation was limited to a small testing set and tracking of a single anatomical landmark.

1.2. Contributions

In this work, a novel motion modelling framework is presented. As shown in Fig. 1, the deep motion model first learns to link complex 3D motion fields with 2D image surrogates through a common latent encoding. The model also learns to recover the 3D motion fields from the latent encoding. On the day of treatment, the proposed framework, composed of a rigid alignment module and the trained deep motion model, is able to process real-time 2D US images of previously unseen cases one by one to provide three-dimensional information about the state of the liver. Once sent to the treatment unit, this information can be used to adjust radiation delivery as needed.

The proposed deep motion model is able to capture a wide variety of motion patterns while also taking into account subject-specific anatomical information to improve its prediction. Our proposed framework does not require prior 3D+t acquisitions for new subjects and removes the need to establish inter-subject correspondences within the training 3D+t data set, an important advantage over previously presented global motion models.

As such, our main contributions are:

- A novel real-time motion modelling framework composed of a rigid alignment module and a deep deformable model evaluated on 20 free-breathing subjects.
- A convolutional autoencoder motion model which learns to recover complex 3D deformations for a previously unseen subject with only a pair of pre-treatment volumes and a single 2D image acquired in real-time.
- The introduction of an image similarity-based rigid alignment strategy to cope with large displacements of the treated organ.

2. Methods

In this section, we present our motion modelling framework. We first formally define the problem at hand. Next, we describe in detail each module composing the proposed motion modelling framework as well as its training procedure. Finally, details of the framework's implementation are provided.

2.1. Problem formulation

We consider a dataset of free-breathing 3D+t US acquisitions of the liver from a population of N individuals. For each subject $s_i \in (s_1, s_2, \dots s_N)$, a sequence of 3D US volumes $\boldsymbol{V} = (V_1, V_2, \dots, V_t)$ is defined, spanning a given time period $[0, t]$. To obtain a temporal sequence of 2D surrogate images $\boldsymbol{I} = (I_1, I_2, \dots, I_t)$, the central slice of each volume $V_t \in \boldsymbol{V}$ is extracted from the chosen anatomical plane. In each sequence $\boldsymbol{V}$, two reference volumes are identified at exhale (V_{exh}) and inhale (V_{inh}) respiratory phases. The rationale for this choice is to cover the entire range of variation during a breathing cycle. The motion observed in the volume sequence can be measured by performing rigid and non-rigid registration between V_{ref} and the current volume $V_t \in \boldsymbol{V}$. Since the exhale phase is a more easily reproducible position for the liver, V_{exh} is chosen as V_{ref}. Hence, the sequence of rigid transformations $\boldsymbol{T} = (T_1, T_2, \dots, T_t)$ and deformation vector fields (DVF) $\boldsymbol{\Phi} = (\phi_1, \phi_2, \dots, \phi_t)$ individually represent the deformations that need to be applied to V_{ref} in order to obtain the corresponding volume V_t. The first step is to compute a 3D rigid transformation between V_{ref} and V_t using a sinle 2D US image I_t, V_{exh} and V_{inh}. Having the rigidly aligned reference volume ($V_t^{rigid} = \mathcal{T}(V_{exh}, T_t)$), the second step is to learn the deformable component to be applied on V_t^{rigid} to match V_t. Therefore, the prediction of each temporal volume is based only on V_{exh}, V_{inh} and I_t as inputs.

2.2. Proposed framework

In the following subsections, we present details about each component of our proposed solution. As shown in Fig. 2, our solution is composed of 2 main components: a rigid alignment module and a deformable motion model, generating 3D volumes in real-time. The rigid alignment module applies an initial rigid displacement to the reference volume in order to coarsely align it with the current position of the liver. The rigidly aligned volume is then fed to the deformable motion model which applies finer localized deformations. The deformable motion model generates its output from a learned low-dimensional encoding of the organ's deformation field and subject-specific features included as skip connections.

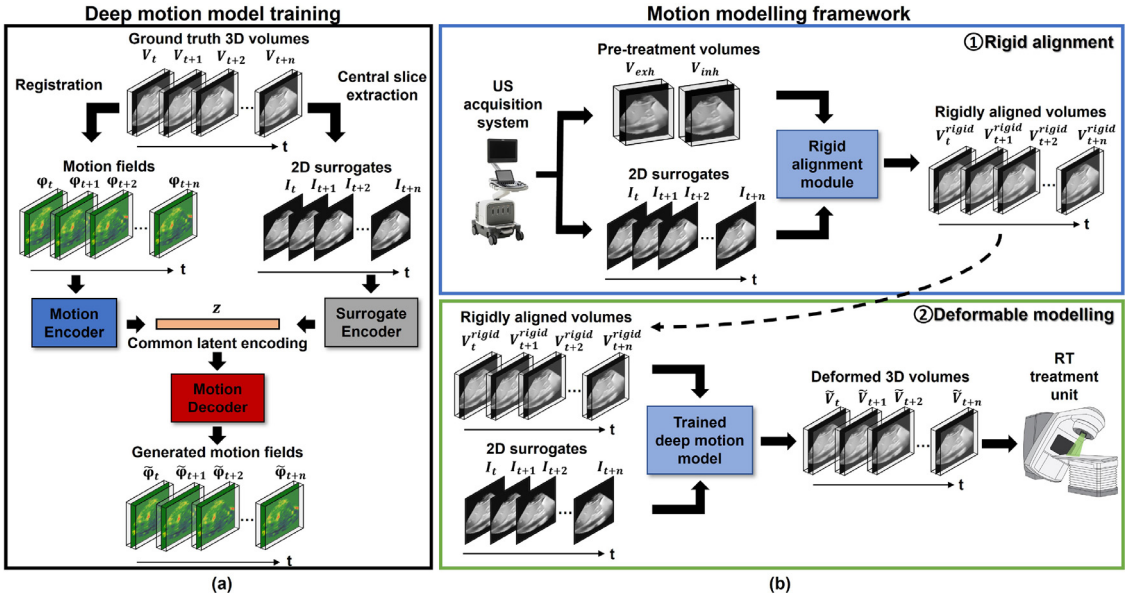


Fig. 1. Overall training and clinical workflow for the proposed respiratory motion modelling framework. (a) A deep respiratory motion model is first trained to associate complex 3D motion fields and 2D surrogate US images through a common latent encoding from a population of subjects. Then, the model is trained to recover the correct input 3D motion fields from the learned latent encoding. (b) On the day of treatment, two pre-treatment volumes (inhale and exhale) are first acquired before treatment for an unseen subject. During treatment, the pre-treatment volumes and real-time 2D images are fed into the proposed framework one by one, generating one deformed 3D volume of the imaged organ per input 2D image. The resulting real-time stream of 3D volumes can be used to adjust the administration of radiotherapy in real-time.

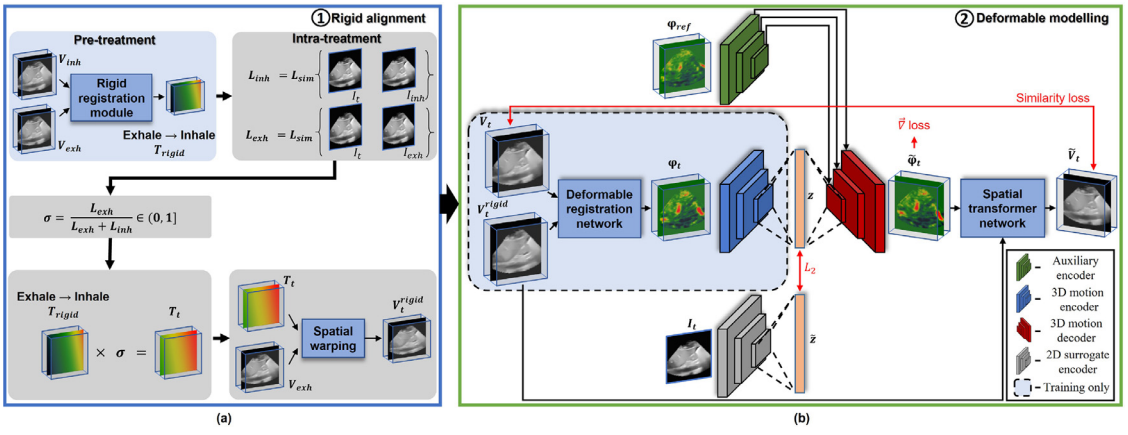


Fig. 2. Schematic representation of the proposed motion modelling framework. (a) First, a rigid transformation is applied to the reference volume in order to coarsely align it with the current state of the liver. The transformation is based on the similarity of the current surrogate 2D image I_t to the central slices of two pre-treatment volumes acquired at exhale I_{exh} and inhale I_{inh}. (b) Once the rigid alignment is performed, the motion autoencoder receives the registration field between V_t and V_t^{rigid} computed by the deformable registration network. The motion field is compressed into the latent vector z and then recovered with the use of prior subject-specific features from the auxiliary encoder. To be able to generate motion fields in the absence of the motion encoder, the 2D surrogate encoder learns to replicate the latent encoding z from the surrogate 2D image. The generated motion field is used to warp V_t^{rigid} through the spatial transformer network (STN) thereby generating the predicted volume $\widehat{V}_t$.

2.2.1. Rigid alignment

Figure 2 a illustrates the proposed approach to rigidly align V_{ref} to V_t during treatment by using two pre-treatment volumes at the extreme respiratory phases and a single 2D US image I_t. Before treatment, two volumes acquired at exhale (V_{exh}) and inhale (V_{inh}) phases are rigidly registered. It is assumed that during treatment, the liver will be located within the exhale-inhale range obtained before treatment. Since the pre-treatment volume at exhale corresponds to V_{ref}, the rigid transformations that will be required to align V_{ref} during treatment are bound between the null transformation and the exhale-inhale transformation. To identify the respiratory phase in which the liver is located during treatment, the current 2D US frame I_t is compared to the corresponding central slices of the pre-treatment volumes I_{exh} and I_{inh} using an image similarity metric $\mathcal{L}_{sim}$. The similarity measures $\mathcal{L}_{exh}$ and $\mathcal{L}_{inh}$ are used to compute a scaling factor σ, as follows:

$$\sigma = \frac{\mathcal{L}_{exh}}{\mathcal{L}_{exh} + \mathcal{L}_{inh}} \in (0, 1]. \tag{1}$$

This factor tends to 0 when I_t is similar to I_{exh} and dissimilar to I_{inh} and tends to 1 in the opposite scenario. In this manner, when the current state of the liver is close to the reference volume (i.e. exhale), a relatively small displacement is applied. As the state of the liver approaches the one in I_{inh}, σ gradually increases and so does the amplitude of the displacement. The obtained value is applied to the exhale-inhale transformation through element-wise multiplication to produce a scaled version which is finally used to generate V_t^{rigid} by applying a rigid transform to V_{ref}. For this work, the Mean Squared Error (MSE) was chosen as $\mathcal{L}_{sim}$, as it was found to be the more efficient when comparing mono-modal images. It is

assumed that I_t, V_{exh} and V_{inh} were all acquired in approximately the same orientation and anatomical location.

2.2.2. Deformable motion modelling

Once rigidly aligned, V_t^{rigid} is fed to the deep deformable motion model shown in Fig. 2b. The goal of this step is to apply a non-rigid 3D deformation ϕ_t to V_t^{rigid} in order to obtain the final 3D output volume $\widetilde{V}_t$ which represents the current state of the imaged organ ($\widetilde{V}_t = \mathcal{T}(V_t^{rigid}, \phi_t)$). First, a pre-trained deformable registration neural network is used to generate the deformation field ϕ_t between V_t^{rigid} and the current volume V_t. In this work, the U-Net like deformable registration network proposed by Balakrishnan et al. (2019) is used for this step. It is important to note that any deep learning based deformable registration network can be used within our framework. Then, the convolutional motion autoencoder is trained to compress each 3D motion field $\phi_t \in \mathbf{\Phi}$ into a corresponding low dimensional latent encoding z. The compression is followed by the recovery of the input motion fields from the obtained latent encoding. Subject-specific information is also incorporated through skip connections which originate from a separate auxiliary encoder. Since the 3D motion fields $\mathbf{\Phi}$ that were used as inputs to the motion autoencoder are not available during inference, a separate 2D surrogate encoder is trained to predict the latent encoding z with a different input. Specifically, the 2D surrogate encoder learns to obtain a latent encoding $\widetilde{z}$ as similar as possible to z by only using a single 2D image of the liver's current state as input. By creating this shared latent representation between the 3D motion autoencoder and the 2D surrogate encoder, the model is capable to infer the complete 3D motion field ϕ_t with only one 2D image as input. To obtain the final predicted volume $\widetilde{V}_t$, a spatial transformation network (STN) warps V_t^{rigid} with the generated 3D motion field $\widetilde{\phi}_t$.

Motion autoencoder The central module of the deformable motion model is the 3D motion autoencoder. Its role is to learn how to compress and recover the input ϕ_t so that during inference, only the latent representation $\widetilde{z}$ is needed to obtain $\widetilde{\phi}_t$. The main components of the autoencoder are the 3D motion encoder and decoder. Both are fully convolutional networks that use strided downsampling operations to reduce the spatial dimension of the input. At the bottleneck of the autoencoder, a latent vector z of size 3072 is obtained by passing the 3D motion encoder's output through one fully connected layer. It is important to note that the dimension of the latent vector z should be determined empirically for this application. An excessively small latent dimension might limit the autoencoder's representational capabilities, while a too large of a latent dimension could lead to an over-parametrization of the network. To recover $\widetilde{\phi}_t$, z is reshaped and passed to the 3D motion decoder. The 3D motion decoder uses transposed convolutions to gradually upsample z back to its original size. Detailed information about the implementation of the network's components is presented in Section 2.3.

Auxiliary autoencoder As the compression of ϕ_t inevitably involves loss of information, the 3D motion decoder bears the complicated task of recreating that lost information using the latent vector z. This is especially challenging when attempted on a previously unseen anatomy during inference. Therefore, to improve the decoder's performance, subject-specific anatomical information is provided at the decoding stage through the use of skip connections (Drozdzal et al., 2016) which carry features from a reference DVF (ϕ_{ref}). Figure 3 shows how ϕ_{ref} is obtained at any time t. First, I_t is replicated along the third dimension to match the size of V_t^{rigid}. This step is required by the deformable registration network which is trained on input volume pairs of the same size. The resulting volume is denoted as V_{rep}. Then, to obtain ϕ_{ref}, the deformable registration network aligns V_t^{rigid} with V_{rep}. Before being included in the decoding process, ϕ_{ref} is processed by the auxiliary encoder which has an identical architecture as the 3D motion encoder. Once encoded, features from each layer of the auxiliary encoder can be concatenated with the features of the analogous decoding layer of the 3D motion decoder. Those skip connections provide the 3D motion decoder with an approximation of the main features of the expected output DVF for the novel unseen anatomy. It is important to note that the number of skip connections can greatly change the behaviour of the network. The optimal number of skip connections to use in this application is determined empirically and shown in Section 3.

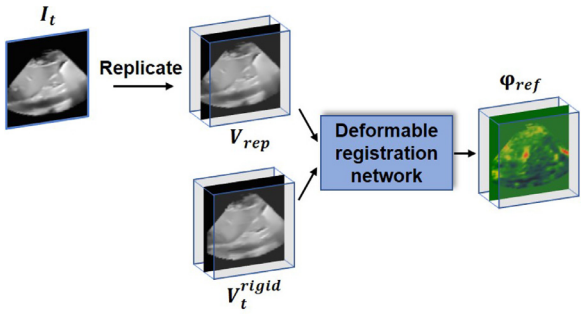


Fig. 3. Schematic representation of the generation of ϕ_{ref}. The surrogate 2D image is replicated to match the dimensions of V_t^{rigid}. Both volumes are then processed by the deformable registration network to obtain an approximation of the DVF to predict.

Surrogate encoder The surrogate encoder aims to regress a latent representation $\widetilde{z}$ as similar to z from a surrogate image I_t. Using this scheme, the 3D deformations are associated with partial observations through their common latent representation. This is achieved by minimizing the following expression:

$$arg\ min\ ||\ z,\ \widetilde{z}\ ||_2^2\ =\ arg\min_{\theta,\omega} ||\ f_\theta(\phi_t),\ g_\omega(I_t)\ ||_2^2 \tag{2}$$

where f_θ and g_ω are functions that parameterize the 3D motion encoder and the surrogate encoder, respectively. The surrogate encoder learns to regress the desired latent encoding z, learned during the autoencoder training, by using the surrogate images provided during treatment. The architecture of the 2D encoder is composed of five 2D convolutional layers, all using strided downsampling layers to reduce the dimension of I_t while gradually increasing the number of channels. Finally, $\widetilde{z}$ is obtained at the output of two fully connected layers. Once the common latent representation is established, the surrogate encoder can replace the 3D motion encoder during inference. In this manner, the full deformation field ϕ_t can be recovered using only the single 2D image I_t as input.

Spatial transformer network The STN module was originally proposed by Jaderberg et al. (2015) to increase the robustness of image registration using convolutional neural networks with respect to spatial variations in their inputs. Since then, it has been used to provide models with the ability to perform spatial warping operations on images and volumes (Balakrishnan et al., 2019; Romaguera et al., 2020). It is comprised entirely of differentiable operations, which is an important property when used in end-to-end trained models. In this work, the STN is used to warp V_t^{rigid} with the predicted deformation fields ϕ_t to obtain the predicted volume $\widetilde{V}_t$, thereby enabling computing the similarity to the true V_t. By using the STN, the motion autoencoder is optimized to predict deformation fields instead of attempting to directly regress the voxel intensities of V_t.

2.2.3. Training procedure and inference

Training The proposed deformable motion model is trained in 3 steps. First, the autoencoder is trained independently, using the 3D

motion fields $\phi_t \in \mathbf{\Phi}$ generated from the registration of V_t^{rigid} and V_t by the deformable registration network. Second, the weights of the autoencoder are fixed while the surrogate encoder is trained to replicate the latent representation of the autoencoder. Finally, all the weights are freed and the entire network is trained together as a final fine-tuning step. During the first step the network is optimized using the first 2 terms of the following loss function:

$$\mathcal{L} = \mathcal{L}_{sim}(\widetilde{V}_t, V_t) + \beta \mathcal{L}_{grad}(\widetilde{\phi}_t) + || z, \widetilde{z} ||_2^2 \quad (3)$$

where the first term ($\mathcal{L}_{sim}$) represents the similarity between the predicted volume $\widetilde{V}_t$ and the true current volume V_t. The second term ($\mathcal{L}_{grad}$), weighted by the parameter β, is a gradient penalty for $\widetilde{\phi}_t$ which encourages the generation of smooth and diffeomorphic deformation fields (Balakrishnan et al., 2019). During the second step, the final loss term in Eq. (3) is used. It represents the L_2 norm between the autoencoder's latent vector z and the surrogate encoder vector $\widetilde{z}$. Finally, in the last step all the terms of Eq. (3) are used to fine-tune all of the network components.

Inference Once trained, the deformable motion model is used without the motion encoder, as shown in Fig. 2b. The inputs during inference are I_t and V_t^{rigid}, which are used to estimate ϕ_{ref}. I_t is used to obtain $\widetilde{z}$, which is passed to the 3D motion decoder. V_t^{rigid} is also used at the last step when it is deformed by the STN to obtain the network's output $\widetilde{V}_t$.

2.3. Implementation details

The proposed model was implemented using PyTorch 1.7.0 (Paszke et al., 2019). The motion encoder is implemented with a 6-layer fully convolutional network. The first three layers include strided downsampling operations with a rate of 2. The number of channels was progressively changed over each layer in the following order $[64, 128, 256, 128, 64, 24]$. The kernel size for all 3D convolutions was $3 \times 3 \times 3$ and the stride and padding were adjusted depending on whether the layer was used for downsampling or not. Each convolutional layer was followed by batch normalization and a ReLU activation layer. The motion decoder is the mirror image of the motion encoder except that all convolutions were replaced by transposed convolutions. Moreover, Leaky ReLU activations with a slope of 0.2 were used for the decoder. The auxiliary encoder has the same architecture as the motion encoder. The surrogate encoder is a fully convolutional network as well. It is comprised of five 2D convolutional layers, four of which use strided downsampling operations. The convolution parameters are the same as for the motion encoder except for the number of channels that was set to $[64, 128, 256, 256, 384]$ to match the dimension of the latent vector. The convolutional layers are followed by two fully-connected layers to regress $\widetilde{z}$.

The Adam optimizer (Kingma and Ba, 2014) was used with an initial learning rate of 10^{-4} which was halved when the validation loss stopped decreasing for 15 epochs. The stopping criteria for step 1 was met when $\mathcal{L}_{sim}(\widetilde{V}_t, V_t)$ did not improve by 0.01 for 10 epochs. The weighting term β in Eq. (3) was set to 0.01. Training in step 2 was stopped when $\mathcal{L}_2(\widetilde{z}, z)$ did not improve by more than 0.01 for 10 epochs. Finally, the stopping criteria for the final training step was the same as step 1 but the threshold was decreased to 10^{-3} to allow for fine-tuning.

For the similarity loss $\mathcal{L}_{sim}$, the MSE loss was slightly adapted for the use with US images. Since US acquisitions appear as a conical shape on a black background, there is a large portion of the voxels that contain no information. A mask representing only non-empty voxels was applied to ignore those regions when registering two volumes or computing image similarity.

A leave-one-out validation scheme was employed to evaluate the network's performance on each unseen subject. Since the deformable registration network is used during inference to generate ϕ_{ref}, it was also trained using the leave-one-out scheme to ensure no data leakage between the model components. Finally, the exhale-inhale transformation used in the proposed rigid alignment module was computed with the widely used medical image registration library Elastix (Klein et al., 2010).

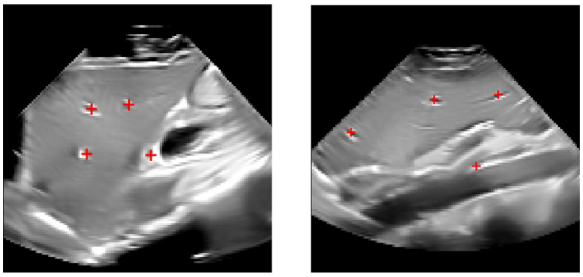
Overall, the reference volume goes through 2 transformations during the execution of the proposed framework. The first transformation (rigid) is done using the warping function from the SimpleElastix package for Python (Marstal et al., 2016). The second warping operation (deformable) is done at the end of the framework. It is performed using the STN module as it allows for end-to-end training. As only two warping steps are performed within our framework, no warping artefacts were observed in the resulting volumes.

3. Experiments and results

In this section, we present the experimental setup used to evaluate the motion modelling framework, with comparisons to state-of-the-art methods. We first present the 3D+t US dataset that was used to train and test the framework. A first set of experiments is presented to analyze the individual contribution of each component to the framework's overall performance. This is achieved through an ablation study and experiments focusing on individual components such as the rigid alignment module, auxiliary encoder and surrogate encoder. Finally, a second set of experiments is conducted to compare our method to other related approaches based on image similarity and target tracking metrics. Results were determined to be statistically different using the Wilcoxon signed rank test with significance level $\alpha = 1\%$. Effect size was measured using Pearson correlation (ρ). Bolded results in tables indicate the best performing model or model configuration for each performance metric. Multiple bolded results for the same metric indicate that there is no significant statistical difference between the results.

3.1. 3D+t US dataset

A dataset of free-breathing 3D+t US sequences was acquired from 20 healthy volunteers, who provided their written consent. The acquisitions were performed using a Philips EPIQ 7G ultrasound system with a X6-1 matrix array transducer. During acquisition, the ultrasound probe was placed under the sternum along the sagittal plane, capturing a cross section of the left liver lobe. The imaging depth was set to 12cm. Focus and contrast were adjusted to provide the best visualization of the liver and its vessels. Using a 15 s acquisition window, up to 3 respiratory cycles were captured with a 250ms temporal resolution, producing sequences of around 60 volumes per volunteer. This yielded the total amount of 1200 volumes in the dataset. The initial average volume size and spatial resolution were $302 \times 228 \times 130$ voxels and $0.58 \times 0.52 \times 0.91$ mm^3 respectively. The acquired volumes were first pre-processed by applying a Bayesian non-local means filter (Coupe et al., 2009) for speckle removal. Then, the volumes were resampled to a 2.0×2.0 mm^2 spatial resolution in the sagittal plane and a slice thickness of 1.0 mm. Finally the volumes were cropped to a size of $64 \times 64 \times 32$ (rows $\times$ columns $\times$ slices). To obtain a sequence of input surrogate images, the central slices of all the volumes composing a given 3D+t sequence were extracted along the desired anatomical plane (sagittal or axial). For each sequence, between 4 and 5 anatomical landmarks were manually annotated by an expert on each temporal volume through one respiratory cycle. As no cancerous tumours were present in the current dataset, the tracked anatomical landmarks were vessels and liver boundaries (see Fig. 4). The estimated inter-rater tracking er-



(a) (b)

Fig. 4. Examples of expert-annotated landmarks placed on sagittal images from the 3D+t US dataset used for evaluation.

Table 1
Resulting image similarity metrics for different model configurations leading to the proposed model. Values are mean ± std.

Model	MSE	NCC	SSIM
Baseline	0.15 ± 0.04	0.42 ± 0.05	0.29 ± 0.05
Baseline + STN	0.10 ± 0.06	0.57 ± 0.10	0.54 ± 0.11
Baseline + STN + ϕ_{ref}	0.07 ± 0.04	0.62 ± 0.10	0.60 ± 0.10
Rigid only	0.10 ± 0.06	0.61 ± 0.11	0.60 ± 0.12
Proposed (axi.)	0.07 ± 0.04	0.63 ± 0.10	0.61 ± 0.10
Proposed (sag.)	**0.06 ±0.03**	**0.66 ±0.09**	**0.65 ±0.08**

ror variability for manual annotations on 3D US sequences is 1.2-1.8 mm on average (Luca et al., 2015).

3.2. Proposed framework analysis

Ablation study In order to better understand the role and contribution of each component of our framework, an ablation study was performed. Different configurations of the proposed deformable model were compared based on the image similarity between ground-truth and predicted volumes. Eq. (4) shows how the similarity between the generated ($\widetilde{V}_t$) and true (V_t) volumes is computed.

$$\text{Similarity}(\widetilde{V}_t, V_t) = \frac{1}{t}\sum_{0}^{t} \mathcal{L}\ (\widetilde{V}_t \circ M_t,\ V_t \circ M_t) \tag{4}$$

Where M_t is the current US mask described in Section 2.3, $\mathcal{L}$ is the test similarity metric and $\circ$ is the Hadamard product. To leverage different ways to evaluate image similarity we used MSE, normalized cross-correlation (NCC) and structural similarity (SSIM) as the test similarity metrics ($\mathcal{L}$).

The baseline version of the model includes an autoencoder and the surrogate encoder without the rigid alignment module. This means that the model attempts to learn how to directly generate volumes by regressing the voxel intensities instead of deformations. To generate deformation fields instead of voxel intensities, the deformable registration network and the STN are added to the baseline. Next, the auxiliary encoder is introduced to assist the model during the decoding stage. Following that, the rigid alignment module from Section 2.2.1 is included upstream to the model, thereby completing all the model components. The proposed model was evaluated when using sagittal and axial orientations for the surrogate image.

Table 1 shows the results of the ablation study, evaluating each configuration based on the similarity metrics. It can be seen that the successive addition of each component allows to improve the output volumes across all similarity metrics. The large improvement from Baseline to Baseline + STN shows that the deformable motion model performs better when it is optimized to generate deformation fields instead of voxel intensities. The addition of the skip connections (extracted from ϕ_{ref}) further improves the output's quality by providing patient-specific information to the decoder. Finally, the addition of the rigidly aligned input gives an additional improvement to the appearance of the output volumes by reducing the amount of motion that needs to be represented by the autoencoder. This shifts the focus of the deformable motion model on more localized motion patterns. Results also show that the model performs better when the sagittal view images are used as surrogate ($\alpha < 0.01$, $\rho > 0.9$). Presumably, this is because the sagittal view covers a larger liver area than the axial view for sequences acquired under the sternum.

Table 2
Displacement (in mm) applied by the rigid alignment module in different respiratory phases with respect to the distance of the chosen inhale volume to the true inhale position. Values are mean ± std.

V_{inh} selection error	Exhale	Mid-cycle	Inhale	Overall
0.0	2.9 ± 0.9	7.0 ± 2.4	12.0 ± 1.3	7.2 ± 3.8
1.5 ± 0.2	2.8 ± 0.9	6.7 ± 2.3	10.5 ± 0.9	6.7 ± 3.3
3.4 ± 0.4	2.5 ± 0.8	6.1 ± 2.0	8.3 ± 0.7	5.8 ± 2.6
5.2 ± 0.7	2.3 ± 0.8	5.3 ± 1.5	6.4 ± 0.4	4.9 ± 2.0
7.4 ± 0.9	2.1 ± 0.7	4.5 ± 1.1	4.6 ± 0.2	4.0 ± 1.5

Rigid alignment module Our next experiment aimed at validating the robustness of the rigid alignment mechanism when the chosen pre-treatment volumes do not represent the full range of motion of the liver during intervention, measuring the robustness towards variation between baseline and online acquisitions.

Generally, V_{exh} can be chosen reliably since the exhale position is easily reproducible. On the other hand, ensuring that V_{inh} represents the deepest breathing amplitude for the entire sequence is not trivial.

In this experiment, we replaced the true V_{inh} by volumes that are adjacent to it in the temporal sequence. We quantified their distance (in mm) to the actual inhale position through rigid registration. Each inhale volume was used by the rigid alignment module to generate a set of rigid transformations covering one respiratory cycle for each case of the data set. Using the same approach as before, the displacement applied by the rigid transformations was computed. The resulting displacement values were then split into 3 respiratory phase groups (exhale, mid-cycle and inhale), each representing 1/3 of the respiratory cycle.

Table 2 shows the displacement introduced by the rigid alignment module at each phase as a function of the average V_{inh} selection error. It can be observed that the overall effect of increasing the selection error induces a decrease in the generated rigid motion amplitude. This effect is most prominent in the phases closest to inhale where the decrease in displacement is almost equal to the shift from the true inhale position. In contrast, volumes at exhale and mid-cycle phases are less affected by the selection error. In summary, the error in the choice of either V_{exh} or V_{inh} has a direct effect on the maximum displacement yielded by the rigid module.

Auxiliary encoder During the motion generation stage, the motion decoder gets information from two sources; the latent vector z and the skip connections from the auxiliary encoder. In order to better understand how the model uses both sources of information, the number of skip connections varied from 1 to 5, starting from the highest resolution layer and going towards the bottleneck of the autoencoder. In essence, allowing for more skip connections means that the model has more information or features from ϕ_{ref}. This can ultimately lead to ignoring completely the information contained in z. To detect when this occurs, the model's autoencoder was tested both with a learned z vector and with a randomly generated vector z_{rand} of the same size as z. Hence, for each configuration, we evaluate whether the information from the latent vector contributes to the model's performance. Figure 5

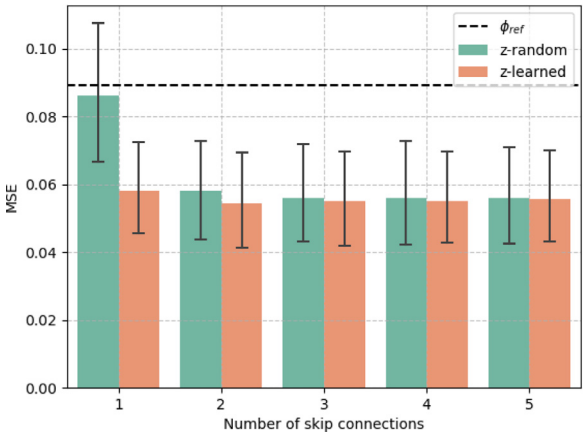


Fig. 5. Motion autoencoder performance with learned and random latent vectors when varying the number of skip connections sent from the auxiliary encoder. The dotted horizontal line indicates the similarity of the volume obtained by directly applying ϕ_{ref} to V_t^{rigid} without going through the autoencoder.

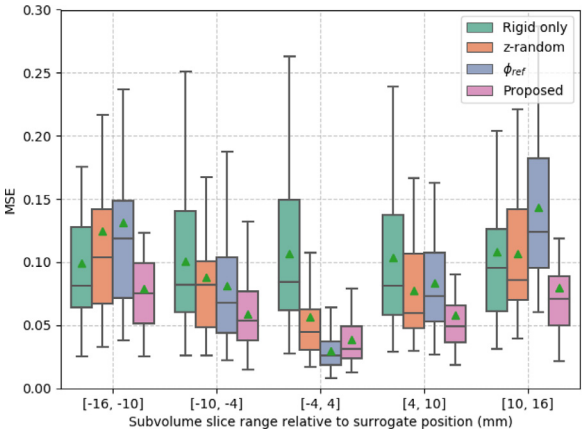


Fig. 6. MSE value distributions between ground-truth and predicted sub-volumes along the medio-lateral axis. The predictions were obtained through either rigid alignment only, deformation with a DVF generated from a random latent vector, deformation using ϕ_{ref} or the proposed model. Mean values are indicated by the green triangles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

presents MSE values between ground-truth and predicted volumes for each model configuration. The dotted horizontal line indicates the similarity of the volume obtained by directly applying ϕ_{ref} to V_t^{rigid} without going through the autoencoder. It is noticeable that for models using 2 or more skip connections, the output's similarity when generated using either z_{rand} or z is practically the same. Indeed, all pairs of results for models with more than one skip connection are statistically the same. This allows us to identify the configuration with one skip connection, at the layer of highest resolution, as the optimal way to introduce patient-specific information from ϕ_{ref}. If more than one skip connection is used, the model tends to ignore the information contained in z and only focuses on the features carried by the skip connections. In that scenario, the model does not take into account any information about the current state of the organ given by I_t.

To further analyze the contribution of the single skip connection and vector z, the image similarity is evaluated at different portions of the output volume. The volumes were split into 5 sub-volumes along the medio-lateral axis. The similarity was evaluated in 4 scenarios: after rigid alignment only (V_t^{rigid}), after warping with ϕ_{ref} only, after warping with a DVF obtained using z_{rand} and after warping with the DVF obtained using the true z vector. Figure 6 shows the MSE between ground-truth and predicted sub-volumes across the different positions. The rigid input volumes V_t^{rigid} show a stable mean similarity across all positions within the complete volume. For volumes warped with ϕ_{ref}, the similarity is better at the center of the volume (i.e near the surrogate image position) and becomes increasingly worse as the sub-volume gets further from the center. This is expected as ϕ_{ref} is generated by registering a volume where I_t is replicated across all slices. Consequently, the most accurate registration is obtained at the center, which is the correct position for I_t. As we move further away from the center, the less accurate the registration becomes. As for volumes warped with a DVF obtained using z_{rand}, a similar conclusion can be made from the skip connections experiment. When one skip connection is used, the model performs worse when provided with random information from the bottleneck. Overall, the best performance was shown by the proposed model that uses the true latent vector z, especially at the edges of the volumes.

Surrogate encoder There is a possibility that the US probe is not positioned at the same location on the patient's body at every fraction of the radiotherapy treatment. Therefore, it is necessary to evaluate the robustness of the model to potential shifts in the position of the 2D surrogate image. To do so, the deformable

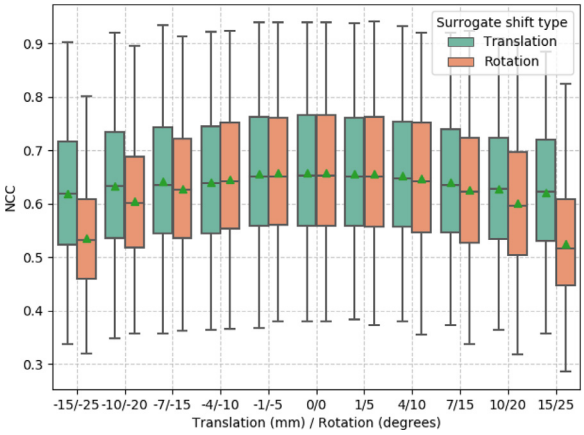


Fig. 7. Image similarity for the entire data set when shifting the position of the surrogate image I_t through translation along the medio-lateral axis or rotation around the sagittal axis.

motion model portion of the framework was first trained to perform predictions based on input 2D images taken from the center of the ground-truth volumes. Then, during inference, the surrogate slice location was changed by either translating it along the medio-lateral axis or by rotating it around the sagittal axis. The translation shift was varied from -15 mm to 15 mm with respect to the central slice, covering the entire volume. The rotation shift was varied from -25 to 25 degrees. Figure 7 shows the NCC between ground-truth and generated volumes across the entire data set when applying both types of surrogate shift independently. It can be observed that, as the translation or rotation shift of the surrogate image increases, the image similarity decreases reaching its minimum at the maximum deviation from the central sagittal slice. The mean difference in NCC does not exceed 0.01 when the shift remains between -4 and 4 mm for translation and -10 to 10 degrees for rotation. Therefore, the model can be considered capable of coping with slight changes in the position and orientation of the surrogate image.

Nevertheless, it is possible to improve the model's tolerance to surrogate shifts and reduce the deterioration of the generated volumes due to deviations from the central sagittal plane. By purposefully providing shifted 2D images to the surrogate encoder during training, it is possible to learn a latent encoding that is less de-

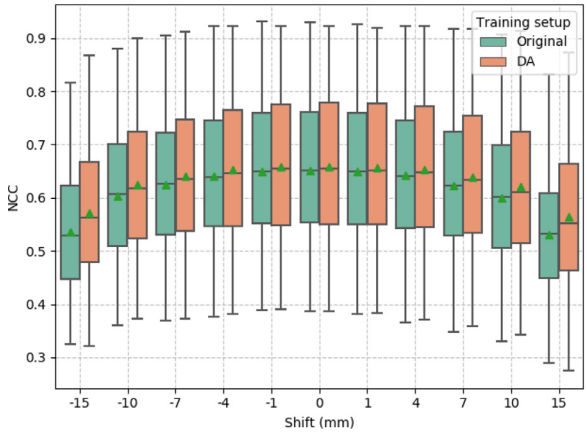


Fig. 8. Evaluation of the robustness to translation shifts in the surrogate image position for the proposed model trained using no DA or DA with shifts of up to 3 mm.

pendent on the position and orientation of the surrogate image I_t. This approach can also be used as a data augmentation (DA) approach to improve model generalization. To evaluate the effect of this training strategy, we re-trained the proposed model by providing surrogate slices with a translation shift of up to 3 mm in each direction along the medio-lateral axis with respect to the central sagittal slice. We excluded rotation shifts and larger translation shifts from this experiment to maintain reasonable training times. Figure 8 shows the performance of the proposed model in its original training setup as well as with the DA strategy. We observe that using the DA strategy helps to improve the proposed model's robustness to shifts in the position of the surrogate image. The largest improvement occurs at more extreme shifts while small shifts do not benefit from this strategy as much.

Real-time application compatibility Finally, to assess the compatibility with real-time applications, the inference time of the proposed framework was evaluated. The total time to process the rigid and deformable steps of the framework was 0.47 ± 0.04 s when executed on CPU and 0.09 ± 0.01 s when executed on a NVIDIA Titan X GPU with 12 GB of RAM. This shows that the required time to generate motion predictions is sufficiently short to be included within a real-time radiotherapy workflow.

3.3. Comparative results

The next set of experiments compared the performance of the proposed framework to related approaches for 3D motion modelling and target tracking in US. Namely, we compare the proposed approach to two other methods described by Paganelli et al. (2018) and Mezheritsky et al. (2020) in the context of image-guided radiation treatments. In the former case, two orthogonal 2D slices extracted from the reference volume (V_{ref}) are registered to the corresponding orthogonal 2D slices in the in-room volume (V_t). Subsequently, the partial 2D motion fields are combined to extrapolate the entire 3D motion. We will refer to this approach as motion extrapolation (ME). As for the model from Mezheritsky et al. (2020), the approach consists of predicting ϕ_t by combining 3D features from V_{ref} and 2D features from I_t. The model is comprised of a 2D encoder for I_t, a 3D encoder for V_{ref} and a 3D decoder coupled with a STN to generate ϕ_t and apply it to V_{ref}. We will refer to this approach as feature combination (FC). All three approaches (ME, FC and the proposed framework) aim to generate the motion field corresponding to the respiratory state indicated by the surrogate 2D information. We first compare their performances based on the similarity metrics used in Section 3.2. We also compute the global and local target registration error (TRE) using 3D deformable image registration (DIR) between ground-truth and predicted volumes, and manual landmark annotations, respectively.

Table 3
Image similarity metrics between ground-truth and predicted volumes for different comparative methods. Values are mean ± std.

Model	MSE	NCC	SSIM
Unregistered	0.09 ± 0.06	0.59 ± 0.11	0.55 ± 0.13
Rigid only	0.10 ± 0.06	0.61 ± 0.11	0.60 ± 0.12
ME (Paganelli et al., 2018)	0.21 ± 0.08	0.59 ± 0.08	0.53 ± 0.10
FC (Mezheritsky et al., 2020)	0.09 ± 0.04	0.57 ± 0.09	0.54 ± 0.10
FC + Rigid	0.08 ± 0.05	0.63 ± 0.10	**0.63 ±0.10**
Proposed	**0.06 ±0.03**	**0.66 ±0.09**	**0.65 ±0.08**

Image similarity Table 3 shows the similarity metrics for the different compared methodologies. As a reference, in the first row of the table, we report the result when there is no motion compensation (Unregistered). The second row represents the values measured when only the rigid alignment is applied on the reference volume. Overall the proposed approach showed the best performance for all metrics except for SSIM where it was statistically equivalent to FC with rigid alignment ($\alpha = 0.4,\ \rho = 0.38$). Although the rigid alignment was designed to be used in the proposed model, it was able to significantly improve the results for the FC model ($\alpha < 0.01, \rho > 0.9$), showing its usability as an independent rigid alignment module. The worst similarity results were obtained by ME, which was designed for local modelling. In consequence, there is a poor overall similarity between ground-truth and predicted volumes.

Target tracking Table 4 compares the methods based on local TRE for different respiratory phases and for the entire respiratory cycle overall. The results were obtained by manually tracking each of the identified landmarks and averaging the difference between the ground-truth and predicted landmark positions. Eq. (5) shows how the TRE is computed from pairs of predicted (p_n^{pred}) and true (p_n^{true}) 3D landmark annotations using the Euclidean distance d.

$$TRE = \frac{1}{n}\sum_{0}^{n} d\ (p_n^{pred},\ p_n^{true}) \tag{5}$$

The results obtained on the reference volume were excluded whenever it was part of the analyzed respiratory cycle. It can be observed that all models were able to improve over the unregistered volumes for all respiratory phases. Moreover, the errors are larger for phases that are temporally further away from the reference respiratory phase. The largest improvement in TRE came from the rigid alignment of the volumes. This is expected as this step aims to represent the general motion of the organ, which includes the largest displacement. Further improvements in the predicted landmark positions result from the proposed deformable modelling. Three models (ME, FC + rigid and proposed) performed similarly at exhale, however as phases get closer to the inhale phase, the proposed approach shows significantly lower errors achieving the best local TRE result overall. In addition, for the inhale phase, the proposed model's highest average tracking error of 1.8 ± 1.5 mm occurs in the SI direction. The average tracking errors in the AP and lateral directions are 1.0 ± 0.9 mm and 0.5 ± 0.6 mm respectively. Since most of the liver's motion occurs in the SI direction, it is expected that the error is the highest in that orientation. Nonetheless, these results show that the proposed model allows to reduce the uncertainty of the target position in the SI direction to a similar order of magnitude as the other directions.

To better visualize the tracking performance of each method, Fig. 9 shows how the tracking error as well as the vessel trajectory in the SI direction evolve over time for each model during 3 respi-

Table 4
3D tracking performance (in mm) of the compared approaches based on local TRE at different phases. Values are mean ± std. ($\mu \pm \sigma$) and 95th percentile (P95).

Model	Exhale		Mid-cycle		Inhale		Overall
	$\mu \pm \sigma$	P_{95}	$\mu \pm \sigma$	P_{95}	$\mu \pm \sigma$	P_{95}	
Unregistered	–	–	9.8 ± 8.2	20.2	18.0 ± 13.4	31.4	10.7 ± 9.7
Rigid only	3.5 ± 1.3	7.6	3.9 ± 1.7	6.9	6.3 ± 4.3	12.6	4.6 ± 3.2
ME (Paganelli et al., 2018)	**2.7 ± 1.4**	**6.1**	5.9 ± 2.8	13.3	10.9 ± 7.9	23.5	6.5 ± 6.4
FC (Mezheritsky et al., 2020)	5.0 ± 3.3	9.7	7.9 ± 4.3	15.4	13.8 ± 10.7	27.5	8.9 ± 7.5
FC + Rigid	**3.1 ± 0.5**	**6.8**	4.5 ± 2.2	6.5	7.2 ± 4.4	10.8	4.9 ± 3.9
Proposed	**2.8 ± 1.6**	**5.6**	**3.2 ± 0.8**	**5.1**	**4.5 ± 2.5**	**9.5**	**3.5 ± 2.4**

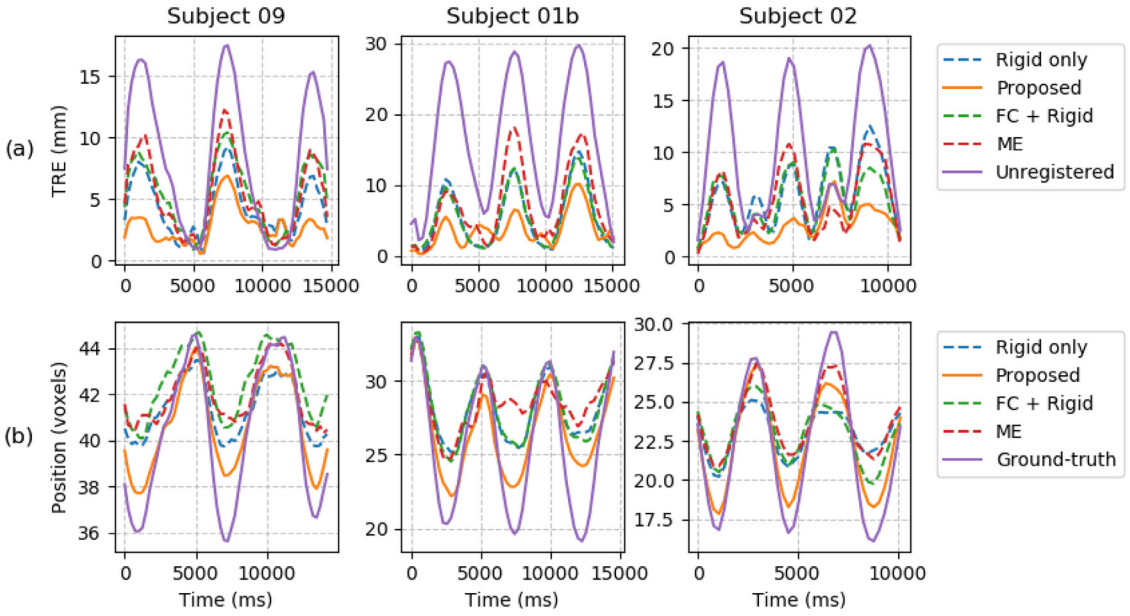


Fig. 9. (a) Evolution of TRE through time and (b) target trajectories in the SI direction for 3 cases. Landmarks were tracked for all 3 acquired breathing cycles.

ratory cycles for 3 subjects within the 3D+t US data set. As previously observed in Table 4, the largest tracking errors occur at the inhale respiratory phase. Figure 9a shows that the proposed framework is able to maintain the lowest error throughout all respiratory cycles compared to other models. The plots in Fig. 9b show that the proposed model is able to follow the ground-truth trajectory better than the comparative approaches.

Furthermore, the proposed model's average tracking error for all cases with respect to the landmark's distance to the central plane of the volume is shown in Fig. 10. It is possible to observe that all values hover around the reported average tracking error for the proposed model regardless of the distance to the central plane. Based on these results, we conclude that it is not the position of the landmark that affects the model's tracking performance, but rather the respiratory phase in which the liver is found.

In addition to local TRE, the global displacement error of the proposed framework was evaluated by applying 3D DIR between the ground-truth and generated volumes. By converting the obtained displacement fields to displacement magnitudes and then averaging over the entire volume, we obtain the average estimation error over all voxels in the generated volume. Eq. (6) shows how the global estimation error is computed by averaging the magnitudes of the displacement field computed between the predicted and ground-truth volumes $\phi(\widetilde{V}_t, V_t)$ using the deformable registration network.

$$\text{Global estimation error} = \frac{1}{t}\sum_{0}^{t} || \ \phi(\widetilde{V}_t, V_t) \ ||_2^2 \tag{6}$$

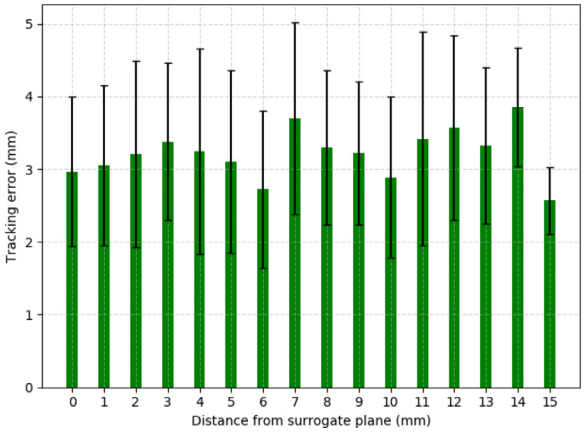


Fig. 10. Average local tracking error (mm) of the proposed model for all cases and all landmarks with respect to the distance between the tracked landmark and the central plane.

The same procedure was applied to the unregistered volumes, however the displacement from the rigid transformation calculated by the rigid module was also taken into account. Figure 11 shows the calculated global estimation error distribution for each case in the US dataset. By observing the different value distributions of the unregistered volumes, it is noticeable that the data set presents a wide variety of motion amplitudes. For all cases, the proposed

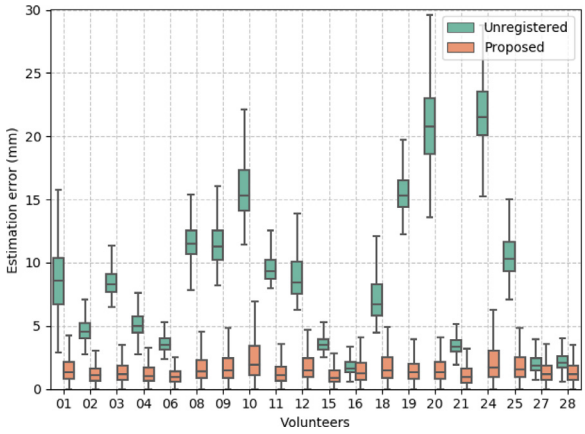


Fig. 11. Global estimation errors calculated by 3D DIR for reference volumes without motion compensation (Unregistered) and volumes generated by the proposed model (Proposed) with respect to the ground-truth volumes. The obtained deformation fields were converted to motion amplitudes (in mm).

model is able to reduce the global estimation error. The mean global error was reduced from 8.7 mm (Unregistered) to 1.7 mm with the proposed solution.

Deformation quality When generating motion from volumetric images, deformation fields are expected to be diffeomorphic to ensure physically plausible displacements. We evaluated the smoothness of the deformation fields produced by our deformable motion model by calculating the Jacobian matrix determinant over the entire motion field ($|J(\widetilde{\phi}_t)|$). The average $|J(\widetilde{\phi}_t)|$ for the entire dataset was 0.97 ± 0.43 with only 1.1% of negative values, indicating that the model produces smooth and plausible deformations with very few foldings.

Qualitative results Fig. 12 illustrates the generated and ground-truth volumes at three respiratory positions (mid-inhale, inhale and mid-exhale) along two imaging planes (sagittal and axial) for one example case. Difference maps with respect to the ground-truth at the inhale phase are presented as well. A general observation is that the proposed framework is able to infer motion outside of the surrogate plane since motion is visible in both perpendicular planes. Furthermore, the difference maps in both planes showcase that the proposed approach generates the lowest voxel intensity errors at inhale. Also, the proposed framework reproduces small features like vessels and liver borders better than the ME and FC approaches as highlighted by the red circles. It is also noticeable that the application of the deformable motion component over the output of the rigid alignment module improves local correspondences with the ground truth volume, highlighting the importance of the second step of the proposed framework.

Finally, Fig. 13 presents generated and ground truth slices at 5 different phases between exhale and inhale. To display the true and generated deformation fields, green and yellow arrows were overlaid on the generated volume slices. Small sections of the deformation fields were increased in size for better visualisation. At the reference phase, the generated deformation field is essentially null. As the phases get closer to inhale, the amplitude of motion applied to the reference volume is gradually increased. It is noticeable that for the majority of positions, the generated motion field follows the expected motion field well. In general the motion is oriented in the inferior direction as expected during inhalation. More localized motion patterns, representing the deformable components of motion, are also present. They can be seen at the left of the images where the heart is visible, as well as in the bottom section of the images. Additional qualitative results can be found in the supplementary materials.

4. Discussion

This work presented a novel 3D motion modelling framework that includes advantages from both population-based and patient-specific models for US-guided radiotherapy procedures. The proposed framework enables target tracking by generating 3D motion fields, which allow to track multiple landmarks simultaneously and in a continuous manner given a reference location for each landmark. In addition, being based on a deep learning model allows reducing the amount of manual preparation and data manipulation required to construct the motion model while also allowing for real-time inference capabilities. The proposed model has demonstrated promising results for image similarity metrics and target tracking when compared to both traditional and deep learning based approaches.

The ablation study has revealed that the inclusion of features through skip connections was the component with greater contribution to the deformable motion modelling component of the framework. By skipping relevant features to the decoder, the model leverages patient-specific information as it attempts to generate a deformation field for an anatomy it has not seen during training. A limitation that is often attributed to population-based models is that by fitting them to a large amount of anatomies, they ultimately learn to represent an average motion field without being able to properly model the motion of each individual subject (McClelland et al., 2013). However, our model is able to avoid this by leveraging the patient-specific features it receives during the motion field generation. Furthermore, the experiments showed the importance of controlling the amount of patient-specific information that is provided to the model. If too many features from ϕ_{ref} are skipped through, the latent representation of the model collapses and no longer captures meaningful information about the current state of the organ. In this case, the model would merely learn to refine ϕ_{ref}. Therefore, the combination of both skip connections with the latent vector z must be optimized to use both sources of information.

When comparing our framework to statistical global model solutions such as in Preiswerk et al. (2014), our framework presents improvements over the motion model construction step. Indeed, when constructing a population model based on PCA, an inevitable step is the establishment of inter-subject correspondences. Usually, this process is done manually or semi-automatically which adds a significant amount of time to the data preparation and model construction. In addition, this limits the type of data that the model can operate on. If the training data doesn't include the inter-subject correspondences established during the model construction, the model's performance will decrease. By employing a deep learning framework for the population based motion model construction, we allow the network to learn those inter-subject correspondences implicitly. It is assumed however that during inference, the provided inputs show the same field-of-view as the ones used during training. However, fine-tuning the model to a new anatomy remains simpler when using a deep network instead of a global statistical model. A final advantage our approach presents over Preiswerk et al. (2014) is the fact that the model analyzes the entire surrogate image as it comes from the acquisition system. This means true 2D surrogate signals are used instead of tracking a fiducial marker within the surrogate images to drive the motion model.

During treatment planning, radio-oncologists add a margin of at least 5 mm around the region to be treated to account for respiratory motion. This is done even in the presence of breath hold techniques (Brock, 2011). Since our model achieved an average TRE of 3.5 mm it could allow to reduce the extent of the added margins, thereby sparing healthy tissues from an unnecessary radiation dose. Although the main focus of this paper is the tracking capa-

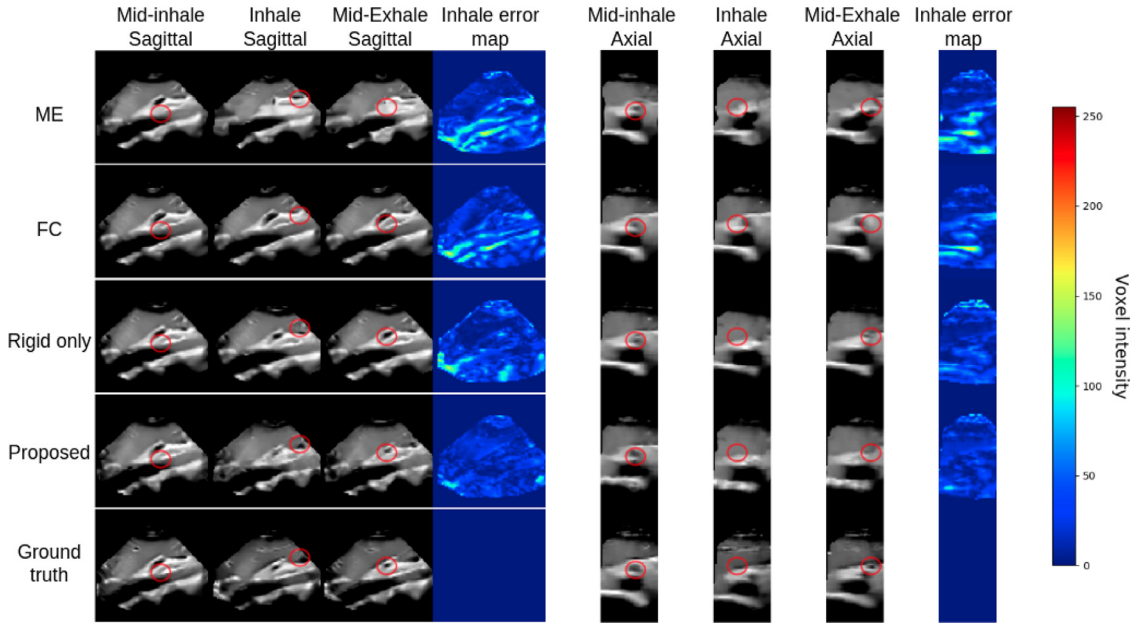


Fig. 12. Qualitative results for all compared methods. For both sagittal and axial planes, the central slice of the volume is shown at mid-inhale, inhale and mid-exhale respiratory phases. For the inhale phase, an error map is calculated and shown. Red circles are included to highlight differences between the displayed approaches.

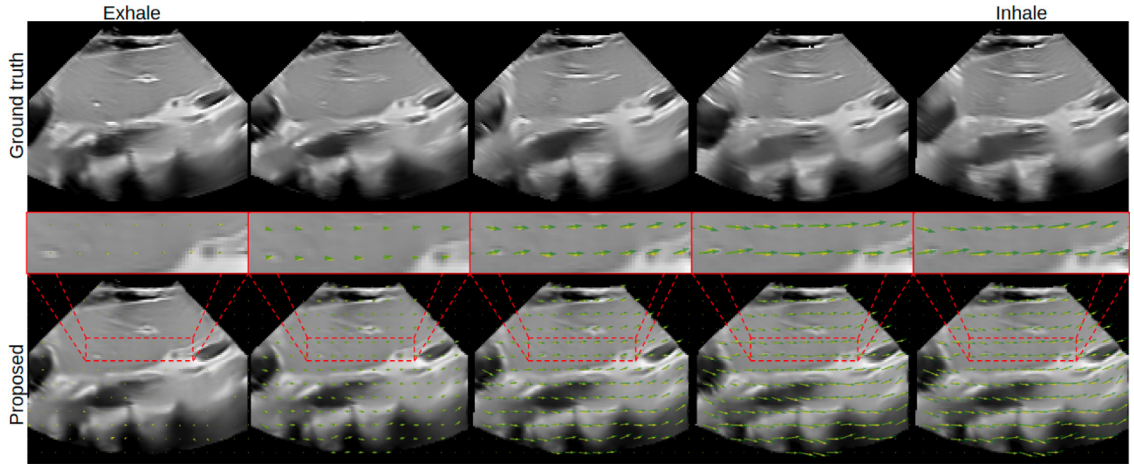


Fig. 13. Qualitative results from exhale to inhale phases with overlaid ground-truth (green) and predicted (yellow) displacement fields. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

bilities of our network, its use is not limited to 3D target tracking. As shown by the image similarity and global TRE experiments, our model can also estimate the global and deformable motions of the organ along all the slices in the volume. This information can be useful for radiotherapy applications, in particular for dose delivery estimation. In addition, by requiring a surrogate image in only one imaging plane, the surrogate acquisition time is reduced without sacrificing the capability of the model to predict the motion outside of that plane.

The local TRE experiment has shown the flexibility of the proposed rigid alignment module as it has improved the performance of 2 different deep learning-based models by providing them with rigidly aligned input volumes. However, this approach is not exempt of limitations. As explained in Section 2.2.1, it is assumed that the state of the liver is bound between the exhale and inhale positions acquired before treatment. In the event where the liver exceeds those bounds, σ no longer describes the position of the liver relative to the pair of pre-treatment volumes, thereby providing less accurate alignment between the central slice of the reference volume and the surrogate image. The first experiment of Section 3.2 has demonstrated that an error in one of the pre-treatment volumes will only affect the accurate alignment of volumes close to the faulty pre-treatment volume. In theory, the deformable motion model could compensate for the lack of displacement of the rigid alignment module in this type of scenario. However, the deformable motion model needs to be trained on volumes that do not present an ideal rigid alignment to provide a sufficient rigid motion compensation. To ensure that the pair of pre-treatment volumes accurately present the full range of motion of the liver, it is recommended that the pre-treatment volumes are acquired before each treatment. This also reduces the negative effects of anatomical changes that occur during the course of the treatment on the rigid alignment module.

Although the proposed rigid alignment module could have been implemented using a deep learning approach, we found that implementing it as an image similarity based module was more advantageous. As rigid alignment only involves the choice of 6 parameters for translation and rotation in 3D, it seemed unnecessary

to increase the complexity of the motion model architecture and training procedure to achieve coarse rigid alignment with an additional deep model. By using the image-similarity based approach, we achieve a sufficient level of performance for the requirements of the deformable model, with very few pre-treatment steps and a very short execution time during treatment.

When developing our proposed framework, we envisioned the following clinical protocol. On the day of treatment, the patient is positioned in the treatment room the same way as during the planning acquisitions, with the US probe in place. Once patient setup is complete, the medical staff acquires the two pre-treatment volumes (exhale and inhale) and begin acquiring the 2D surrogate signal. At this time, the motion modelling framework can be initialized and used to communicate with the treatment unit. As the probe cannot be manipulated by medical staff during the treatment, it will need to be held by a dedicated probe-holding device. For instance, it can be a robotic arm such as in Schlüter et al. (2019) or a 3D printed in-house tool. The choice of the probe-holding device will depend on the position at which the probe will need to be held.

An important feature that motion models need to have is the ability to predict and anticipate the motion the target will experience in real time. This is necessary because the adjustment of the treatment plan and delivery to a new position of the target isn't instantaneous and bears a latency that cannot be ruled out (Keall et al., 2006). Several works on motion modelling have presented ways to include motion prediction within their framework (Preiswerk et al., 2014; Harris et al., 2016; Romaguera et al., 2020). While in this work the model doesn't present motion prediction capabilities, the framework is capable of including a temporal prediction module. Specifically, the surrogate branch of the motion model can learn to predict the future latent representation of the organ, thus generating the future anticipated motion field. While this addition is crucial for the applicability in a clinical setting, it is out of the scope of this work and needs to be validated in future studies.

Another limitation common to several motion models which hinders the transfer of those approaches to the treatment room, is the amount of subjects used for validation. In this work, the data acquired from the 20 subjects presented a good variability in anatomical appearance. However, the acquisition time for each sequence (15 s) has limited the amount of breathing variability that was captured. Also, long-term effects such as exhale drift (von Siebenthal et al., 2007) could not be taken into account either. Since the 3D+t dataset was acquired on healthy subjects only, the proposed solution was not evaluated on liver cancer patients undergoing radiotherapy treatment. As those cases can present higher variability in anatomical appearance and breathing patterns, due to the presence of tumors or other pathologies, the robustness of our framework needs to be validated on this type of data in future studies. Moreover, as explained in Section 2.1, this study assumes that the reference volumes used in our experiments are directly taken from the acquired 3D+t sequences. The surrogate 2D images are also assumed to be the central slices of the volumes within the 3D+t sequences. In a clinical setting, this wouldn't be the case as there would be no prior 3D+t acquisition. However, we do not believe the framework's performance will be affected as long as the V_{ref} and surrogate image show the same anatomical location and field-of-view. This aspect would need to be validated in future experiments.

Future studies will include the addition of a temporal prediction mechanism, thus increasing the horizon for temporal sequences, with the evaluation on longer sequences, application for different imaging modalities as well as general improvements to individual components such as the rigid alignment module and motion modelling network. A prospective study with radiotherapy patients is planned to further evaluate in a clinical context.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Tal Mezheritsky: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing – review & editing. **Liset Vázquez Romaguera:** Conceptualization, Methodology, Software, Investigation, Visualization, Writing – review & editing. **William Le:** Writing – review & editing. **Samuel Kadoury:** Conceptualization, Resources, Writing – review & editing, Project administration, Funding acquisition.

Acknowledgments

We thank Charlotte Rémy from the department of radio-oncology of the CHUM for helping with the data acquisition. This work was partly funded by an NSERC collaborative research and development project (CRDPJ-517413-17) in collaboration with Elekta Ltd., by the TransMedTech Institute and the FRQNT (OncoTech). We also thank Hongliang Li from the CHUM research center for help with data preprocessing.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2021.102260.

References

Arnold, P., Preiswerk, F., Fasel, B., Salomir, R., Scheffler, K., Cattin, P.C., 2011. 3D organ motion prediction for MR-guided high intensity focused ultrasound. In: Fichtinger, G., Martel, A., Peters, T. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 623–630.

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

Banerjee, J., Klink, C., Vast, E., Niessen, W., Moelker, A., van Walsum, T., 2015. A combined tracking and registration approach for tracking anatomical landmarks in 4D ultrasound of the liver. In: MICCAI Workshop: Challenge on Liver Ultrasound Tracking, pp. 36–43.

Baumann, M., Krause, M., Hill, R., 2008. Exploring the role of cancer stem cells in radioresistance. Nat. Rev. Cancer 8 (7). doi:10.1038/nrc2419.

Boye, D., Samei, G., Schmidt, J., Székely, G., Tanner, C., 2013. Population based modeling of respiratory lung motion and prediction from partial information. In: Medical Imaging 2013: Image Processing, vol.~8669. International Society for Optics and Photonics, p. 86690U.

Brock, K.K., 2011. Imaging and image-guided radiation therapy in liver cancer. Semin. Radiat. Oncol. 21 (4), 247–255. doi:10.1016/j.semradonc.2011.05.001. Radiation Therapy of Primary and Metastatic Liver Tumors

Brock, K.K., Dawson, L.A., 2010. Adaptive management of liver cancer radiotherapy. Semin. Radiat. Oncol. 20 (2), 107–115. doi:10.1016/j.semradonc.2009.11.004. Adaptive Radiotherapy

Coupe, P., Hellier, P., Kervrann, C., Barillot, C., 2009. Nonlocal means-based speckle filtering for ultrasound images. IEEE Trans. Image Process. 18 (10), 2221–2229. doi:10.1109/TIP.2009.2024064.

Davies, S.C., Hill, A.L., Holmes, R.B., Halliwell, M., Jackson, P.C., 1994. Ultrasound quantitation of respiratory organ motion in the upper abdomen. Br. J. Radiol. 67 (803), 1096–1102. doi:10.1259/0007-1285-67-803-1096. PMID: 7820402

Dormand, E.-L., Banwell, P.E., Goodacre, T.E., 2005. Radiotherapy and wound healing. Int. Wound J. 2 (2), 112–127. doi:10.1111/j.1742-4801.2005.00079.x.

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C., 2016. The importance of skip connections in biomedical image segmentation. In: Deep Learning and Data Labeling for Medical Applications. Springer, pp. 179–187.

Fontanarosa, D., van der Meer, S., Bamber, J., et al., 2015. Review of ultrasound image guidance in external beam radiotherapy: I. Treatment planning and inter-fraction motion management. Phys. Med. Biol. 60 (3), R77–R114. doi:10.1088/0031-9155/60/3/r77.

Giger, A., Sandkühler, R., Jud, C., Bauman, G., Bieri, O., Salomir, R., Cattin, P., 2018. Respiratory motion modelling using cGANs. MICCAI.
Gillies, D.J., Gardi, L., De Silva, T., Zhao, S.-r., Fenster, A., 2017. Real-time registration of 3D to 2D ultrasound images for image-guided prostate biopsy. Med. Phys. 44 (9), 4708–4723. doi:10.1002/mp.12441.
Ha, I.Y., Wilms, M., Handels, H., Heinrich, M.P., 2019. Model-based sparse-to-dense image registration for realtime respiratory motion estimation in image-guided interventions. IEEE Trans. Biomed. Eng. 66 (2), 302–310. doi:10.1109/TBME.2018.2837387.
Harris, W., Ren, L., Cai, J., Zhang, Y., Chang, Z., Yin, F.-F., 2016. A technique for generating volumetric cine MRI (VC-MRI). Int. J. Radiat.Oncol.*Biol.*Phys. 95. doi:10.1016/j.ijrobp.2016.02.011.
Hawkes, D., Barratt, D., Blackall, J., Chan, C., Edwards, P., Rhode, K., Penney, G., McClelland, J., Hill, D., 2005. Tissue deformation and shape models in image-guided interventions: a discussion paper. Med. Image Anal. 9 (2), 163–175. doi:10.1016/j.media.2004.11.007. Medical Simulation - Delingette
He, J., Shen, C., Huang, Y., Wu, J., 2019. Siamese spatial pyramid matching network with location prior for anatomical landmark tracking in 3-dimension ultrasound sequence. In: Lin, Z., Wang, L., Yang, J., Shi, G., Tan, T., Zheng, N., Chen, X., Zhang, Y. (Eds.), Pattern Recognition and Computer Vision. Springer International Publishing, Cham, pp. 341–353.
Huang, P., Yu, G., Lu, H., Liu, D., Xing, L., Yin, Y., Kovalchuk, N., Xing, L., Li, D., 2019. Attention-aware fully convolutional neural network with convolutional long short-term memory network for ultrasound-based motion tracking. Med. Phys. 46 (5), 2275–2285. doi:10.1002/mp.13510.
Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks. In: Advances in Neural Information Processing Systems, pp. 2017–2025.
Jaffray, D.A., 2015. Radiation therapy for cancer, pp. 239–247. chapter 14
Jud, C., Cattin, P.C., Preiswerk, F., 2017. Chapter 14 - statistical respiratory models for motion estimation. In: Zheng, G., Li, S., Székely, G. (Eds.), Statistical Shape and Deformation Analysis. Academic Press, pp. 379–407. doi:10.1016/B978-0-12-810493-4.00017-1.
Keall, P.J., Mageras, G.S., Balter, J.M., Emery, R.S., Forster, K.M., Jiang, S.B., Kapatoes, J.M., Low, D.A., Murphy, M.J., Murray, B.R., Ramsey, C.R., Van Herk, M.B., Vedam, S.S., Wong, J.W., Yorke, E., 2006. The management of respiratory motion in radiation oncology report of aapm task group 76a). Med. Phys. 33 (10), 3874–3900. doi:10.1118/1.2349696.
King, A., Buerger, C., Tsoumpas, C., Marsden, P., Schaeffter, T., 2012. Thoracic respiratory motion estimation from MRI using a statistical model and a 2-D image navigator. Med. Image Anal. 16 (1), 252–264. doi:10.1016/j.media.2011.08.003.
Kingma, D. P., Ba, J., 2014. Adam: a method for stochastic optimization. 1412.6980.
Klein, S., Staring, M., Murphy, K., A. Viergever, M., Josien, P.P., 2010. elastix: A toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging 29 (1), 196–205.
Liu, F., Liu, D., Tian, J., Xie, X., Yang, X., Wang, K., 2020. Cascaded one-shot deformable convolutional neural networks: developing a deep learning model for respiratory motion estimation in ultrasound sequences. Med. Image Anal. 65, 101793. doi:10.1016/j.media.2020.101793.
Luca, V.D., Benz, T., Kondo, S., et al., 2015. The 2014 liver ultrasound tracking benchmark. Phys. Med. Biol. 60 (14), 5571–5599. doi:10.1088/0031-9155/60/14/5571.
Marstal, K., Berendsen, F., Staring, M., Klein, S., 2016. SimpleElastix: a user-friendly, multi-lingual library for medical image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.
McClelland, J., 2013. Estimating Internal Respiratory Motion from Respiratory Surrogate Signals Using Correspondence Models. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 187–213. doi:10.1007/978-3-642-36441-9_9.
McClelland, J., Hawkes, D., Schaeffter, T., King, A., 2013. Respiratory motion models: a review. Med. Image Anal. 17 (1), 19–42. doi:10.1016/j.media.2012.09.005.
McClelland, J.R., Modat, M., Arridge, S., Grimes, H., D'Souza, D., Thomas, D., O'Connell, D., Low, D.A., Kaza, E., Collins, D.J., et al., 2017. A generalized framework unifying image registration and respiratory motion models and incorporating image reconstruction, for partial image data or full images. Phys. Med. Biol. 62 (11), 4273.
Mezheritsky, T., Romaguera, L.V., Kadoury, S., 2020. 3D ultrasound generation from partial 2D observations using fully convolutional and spatial transformation networks. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1808–1811. doi:10.1109/ISBI45749.2020.9098423.
Noorda, Y., Bartels, L., Viergever, M., Pluim, J., 2016. Subject-specific four-dimensional liver motion modeling based on registration of dynamic MRI. J. Med. Imaging 3, 015002. doi:10.1117/1.JMI.3.1.015002.
Ozkan, E., Tanner, C., Kastelic, M., Mattausch, O., Makhinya, M., Goksel, O., 2017. Robust motion tracking in liver from 2D ultrasound images using supporters. Int. J. Comput. Assist. Radiol. Surg. 12. doi:10.1007/s11548-017-1559-8.
Paganelli, C., Lee, D., Kipritidis, J., Whelan, B., Greer, P., Baroni, G., Riboldi, M., Keall, P., 2018. Feasibility study on 3D image reconstruction from 2D orthogonal cine-MRI for MRI-guided radiotherapy. J. Med. Imaging Radiat. Oncol. 62. doi:10.1111/1754-9485.12713.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: an imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp. 8024–8035.
Pham, J., Harris, W., Sun, W., Yang, Z., Yin, F.-F., Ren, L., 2019. Predicting real-time 3D deformation field maps (DFM) based on volumetric cine MRI (VC-MRI) and artificial neural networks for on-board 4D target tracking: a feasibility study. Phys. Med. Biol. 64. doi:10.1088/1361-6560/ab359a.
Preiswerk, F., De Luca, V., Arnold, P., Celicanin, Z., Petrusca, L., Tanner, C., Bieri, O., Salomir, R., Cattin, P., 2014. Model-guided respiratory organ motion prediction of the liver from 2D ultrasound. Med. Image Anal. 18. doi:10.1016/j.media.2014.03.006.
Romaguera, L.V., Plantefève, R., Romero, F.P., Hébert, F., Carrier, J.-F., Kadoury, S., 2020. Prediction of in-plane organ deformation during free-breathing radiotherapy via discriminative spatial transformer networks. Med. Image Anal. 64, 101754. doi:10.1016/j.media.2020.101754.
Royer, L., Krupa, A., Dardenne, G., Le Bras, A., Marchand, E., Marchal, M., 2017. Real-time target tracking of soft tissues in 3D ultrasound images based on robust visual information and mechanical simulation. Med. Image Anal. 35, 582–598. doi:10.1016/j.media.2016.09.004.
Samei, G., Goksel, O., Lobo, J., Mohareri, O., Black, P., Rohling, R., Salcudean, S., 2018. Real-time FEM-based registration of 3-D to 2.5-D transrectal ultrasound images. IEEE Trans. Med. Imaging 37 (8), 1877–1886. doi:10.1109/TMI.2018.2810778.
Samei, G., Tanner, C., Székely, G., 2012. Predicting liver motion using exemplar models, pp. 147–157. doi:10.1007/978-3-642-33612-6_16.
Sawada, A., Yoda, K., Kokubo, M., Kunieda, T., Nagata, Y., Hiraoka, M., 2004. A technique for noninvasive respiratory gated radiation treatment system based on a real time 3d ultrasound image correlation: a phantom study. Med. Phys. 31 (2), 245–250. doi:10.1118/1.1634482.
Schlüter, M., Fürweger, C., Schlaefer, A., 2019. Optimizing robot motion for robotic ultrasound-guided radiation therapy. Phys. Med. Biol. 64 (19), 195012. doi:10.1088/1361-6560/ab3bfb.
Schweikard, A., Glosser, G., Bodduluri, M., Murphy, M.J., Adler, J.R., 2000. Robotic motion compensation for respiratory movement during radiosurgery. Comput. Aided Surg. 5 (4), 263–277. doi:10.3109/10929080009148894.
Selmi, S.-Y., Promayon, E., Troccaz, J., 2018. Hybrid 2D-3D ultrasound registration for navigated prostate biopsy. Int. J. Comput. Assist. Radiol. Surg. 13, 987–995.
Shepard, A., Wang, B., Foo, T., Bednarz, B., 2017. A block matching based approach with multiple simultaneous templates for the real-time 2D ultrasound tracking of liver vessels. Med. Phys. 44. doi:10.1002/mp.12574.
Stemkens, B., Tijssen, R.H., De Senneville, B.D., Lagendijk, J.J., Van Den Berg, C.A., 2016. Image-driven, model-based 3d abdominal motion estimation for mr-guided radiotherapy. Phys. Med. Biol. 61 (14), 5335.
Suramo, I., Päivänsalo, M., Myllylä, V., 1984. Cranio-caudal movements of the liver, pancreas and kidneys in respiration. Acta Radiologica Diagnosis 25 (2), 129–131. doi:10.1177/028418518402500208. PMID: 6731017
Tanner, C., Zur, Y., French, K., Samei, G., Strehlow, J., Sat, G., Donald-Simpson, H., Houston, J., Kozerke, S., Székely, G., Melzer, A., Preusser, T., 2016. In vivo validation of spatio-temporal liver motion prediction from motion tracked on MR thermometry images. Int. J. Comput. Assist. Radiol. Surg. 11. doi:10.1007/s11548-016-1405-4.
von Siebenthal, M., Szekely, G., Gamper, U., Boesiger, P., Lomax, A., Cattin, P., 2007. 4D MR imaging of respiratory organ motion and its variability. Phys. Med. Biol. 52 (6), 1547.
Weiss, P.H., Baker, J.M., Potchen, E.J., 1972. Assessment of hepatic respiratory excursion. J. Nucl. Med. 13 (10), 758–759.
Western, C., Hristov, D., Schlosser, J., 2015. Ultrasound imaging in radiation therapy: from interfractional to intrafractional guidance. Cureus 7 (6). doi:10.7759/cureus.280.