A deep learning framework for 2D-3D image registration and volumetric imaging in the presence of biologically-driven motion

Nicholas Hindley (nicholas.hindley@sydney.edu.au)
Harvard University
Paul Keall
University of Sydney
Chun-Chien Shieh
University of Sydney

Article

Keywords:

Posted Date: June 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1741952/v1

Additional Declarations: Competing interest reported. N.H., A.S. and P.K. are inventors on a filed provisional patent for the method disclosed in this paper.

A deep learning framework for 2D-3D image registration and volumetric imaging in the presence of biologically-driven motion

Nicholas Hindley[1,2], Chun-Chien Shieh[1,3], Paul Keall[1]

[1]ACRF Image X Institute, University of Sydney, Sydney, NSW, Australia

[2]A. A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

[3] Sydney Neuroimaging Analysis Centre, University of Sydney, Sydney, Australia

Biomedical imaging often relies on 2D images to capture information about moving 3D objects. However, without sufficient prior knowledge, the problem of mapping from 2D images to 3D motion is computationally intractable. In this paper we introduce *Voxelmap*, a deep learning framework that achieves 2D-3D image registration and volumetric imaging in real-time by imposing implicit constraints on biologically-driven motion. Here we demonstrate the use of this framework in image-guided radiotherapy with data from two lung cancer patients. By efficiently estimating biologically-driven 3D motion from 2D images, this framework could also find application in contexts such as fetal imaging and functional neuroimaging.

Many scenarios in biomedical imaging involve acquiring 2D images in the presence of 3D motion. Examples of such scenarios include functional neuroimaging[1-3], where changes in blood flow are used to assess neuronal activity; fetal imaging[4,5], where detailed images of the heart and brain are used to diagnose congenital disease; and image-guided radiotherapy[6,7], where radiation beams must adapt to the motion of tumors and organs-at-risk. In these scenarios, motion estimation could be used to achieve real-time adaptation or to improve image reconstruction, but the task of mapping from 2D images to 3D motion is severely ill-posed. The changing position and orientation of the imaging device with respect to the moving objects may not be known, the acquired images can be corrupted by noise, scatter or motion blur, and there are often several ways of orienting a 2D image that fit the 3D coordinate space equally well[8]. This problem of determining an appropriate transformation between a 2D image and a 3D image given a shared coordinate system is known as 2D-3D image registration. Without sufficient prior knowledge, the solution space for 2D-3D image registration is arbitrarily large, making efficient computation intractable. In this paper, we propose a deep learning framework to estimate biologically-driven 3D motion from 2D images. Here we demonstrate the use of this framework to achieve real-time respiratory motion estimation and volumetric imaging from images acquired during lung cancer radiotherapy.

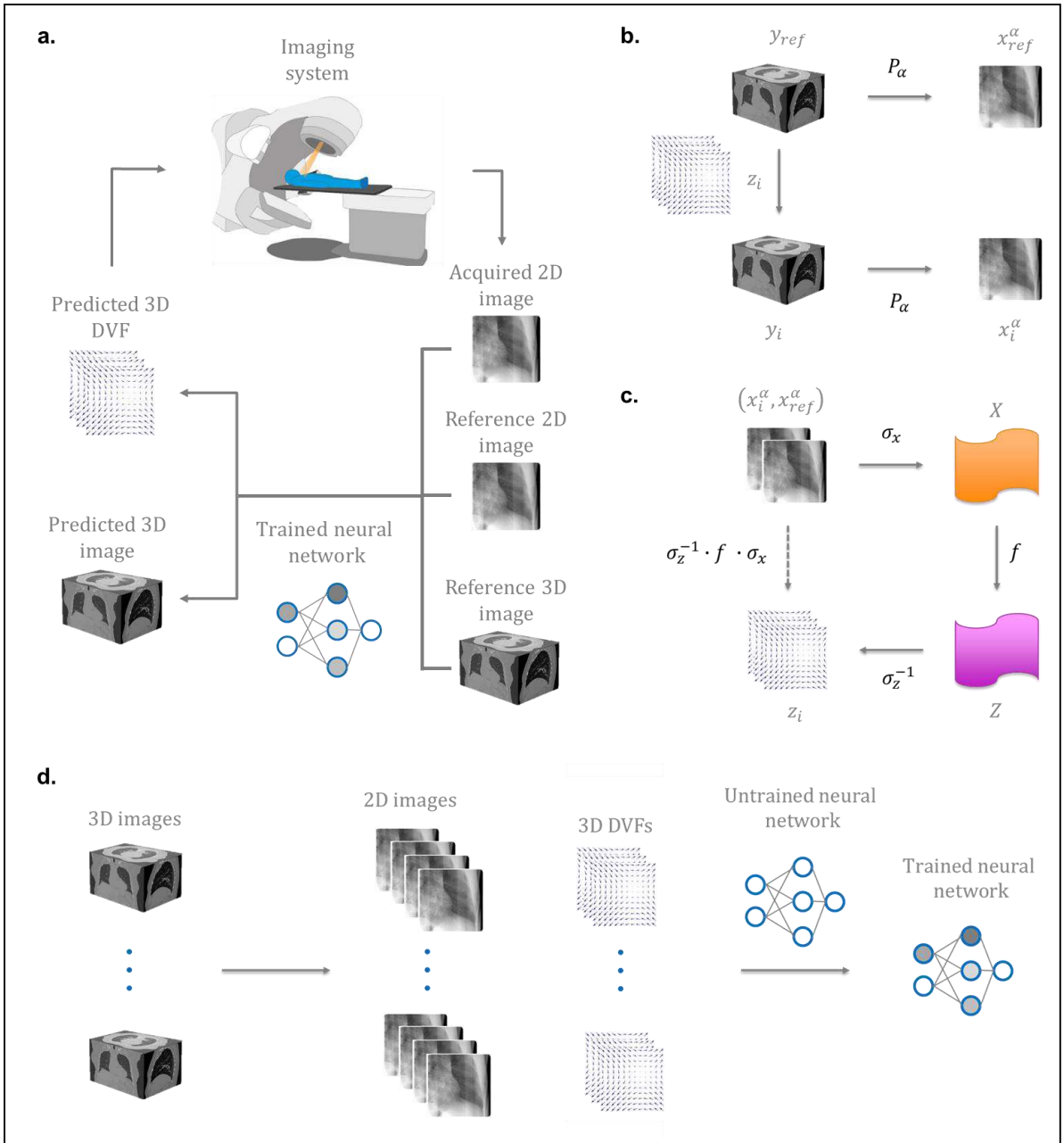

Fig 1. | Real-time 2D-3D image registration and volumetric imaging with Voxelmap a, During online imaging, a trained neural network takes an acquired 2D image from the imaging system as input in addition to a reference 2D image and reference 3D image. The network then predicts a 3D deformation vector field (DVF) that can be fed back into the imaging system for real-time adaptation. Additionally, this vector field can be used to warp the reference 3D image to produce a predicted 3D image for volumetric imaging. **b**, In 2D-3D image registration, certain "allowed" patterns of 3D motion exist given the corresponding 2D views. 3D images y_i and y_{ref} are related by the 3D DVF z_i, while 2D images x_i^{α} and x_{ref}^{α} are obtained by applying the projection operator P_α to y_i and y_{ref} respectively. **c**, Mapping from 2D images to 3D DVFs can be considered as a problem of manifold approximation. More specifically, σ_x maps 2D image pair $\left(x_i^{\alpha}, x_{ref}^{\alpha}\right)$ to the manifold over 2D image pairs X, f maps from X to the manifold over 3D DVFs Z and σ_z^{-1} maps from Z to the candidate 3D vector field z_i. A key hypothesis of this paper is that the composite transformation $\sigma_z^{-1} \cdot f \cdot \sigma_x$ can be learned by a deep neural network. **d**, Prior to online imaging, 3D images are used to produce 2D images and 3D DVFs for network training.

Approximately half of all cancer patients can benefit from some form of radiotherapy[9-11], but the commonest causes of cancer-related death involve tumors in the thorax and abdomen[12] that are constantly moving due to respiration[13]. In the presence of this motion, radiation beams must be enlarged to encompass tumour position over all points in the respiratory cycle but this comes at the cost of healthy tissue damage. Real-time motion management can be used to maintain minimal target volumes in the presence of respiration but these technologies require dedicated systems that are not available in most radiotherapy centers[14-20]. A recent survey of 200 centers across 41 countries found that 71% wished to extend real-time motion management to additional treatment sites, but were hindered by human and financial resources as well as machine capacity[21]. This indicates a strong demand for inexpensive motion management strategies that can be seamlessly integrated into the existing clinical workflow without the need for additional equipment. Additionally, there is currently no technology that provides real-time 3D visualisation of changing patient anatomy during diagnostic or interventional procedures. In this paper, we introduce *Voxelmap*, a deep learning framework for 2D-3D image registration and volumetric imaging (Fig 1a). The core idea behind our proposed approach is that 3D motion estimation and volumetric imaging can be made computationally tractable by considering that 2D views provide hints about 3D biologically-driven motion. For instance, inhalation involves simultaneous inferior and anterior motion of the diaphragm such that an inferior shift in the coronal plane corresponds to an anterior shift in the sagittal plane, lateral expansion of the heart in the coronal plane corresponds to dorsiventral expansion in the axial plane and the trajectories of cerebral blood flow follow detailed paths sketched out by veins and arteries. In other words, certain "allowed" patterns of 3D motion exist given the corresponding 2D views (Fig 1b). Once this 3D motion is estimated it then can be used to warp a static 3D image, enabling volumetric imaging.

Inspired by image reconstruction by domain-transform manifold learning[22], we consider the problem of mapping from the 2D images to 3D motion as one of manifold approximation (Fig 1c). A *manifold* is a locally Euclidean topological space. For instance, when navigating the surface of the Earth one uses an atlas that considers each local neighbourhood of points as being situated on flat map, despite global curvature. Considering navigation in these terms is useful because Euclidean space allows for the use of convenient mathematical tools including, in the case of differentiable manifolds, calculus. Combining the mathematics of manifolds with the intuition that 3D biologically-driven motion is constrained by 2D views we formulate the problem of respiratory motion estimation during image-guided radiotherapy as follows. The standard clinical workflow for thoracoabdominal targets begins by acquiring a 4D-CT[23-25], which is set of 10 CT images $(y_1, \dots, y_{10})$ where each image represents the 3D internal anatomy of a patient over the course of respiration. For example, y_1 corresponds to the 3D image at peak-inhalation, y_2 corresponds to the 3D image at 20 % exhalation, et cetera. The peak-exhalation image is typically selected as the reference image y_{ref} for segmentation and treatment planning as it usually contains the fewest motion artefacts. Image registration to y_{ref} yields a set of 3D deformation vector fields (DVFs) $(z_1, \dots, z_{10})$ which represent how every voxel in y_{ref} moves from each image in the set $(y_1, \dots, y_{10})$. Additionally, since the angles at which the patient will be imaged and irradiated on the treatment day are often known beforehand, we can forward-project each 3D image y_i at n angles to obtain the set of 2D images $\left(x_i^1, \dots, x_i^n\right)$. We wish to know how the positions of the target and organs-at-risk segmented on the planning day vary during treatment. To achieve this task we consider mapping from 2D image pair $\left(x_i^\alpha, x_{ref}^\alpha\right)$ acquired as projections through 3D image pair $\left(y_{i,} y_{ref}\right)$ at angle α to the 3D DVF z_i. Importantly, there are certain key anatomical features, such as the diaphragm and intercostal muscles, that are visible in each 2D image pair and that drive motion in each 3D DVF. Additionally, since every 2D image and 3D DVF differs only by respiratory motion for

a given patient, this set of key features can be embedded in an abstract topological space that varies smoothly. In other words, there exists a function σ_x that maps $\left(x_i^\alpha, x_{ref}^\alpha\right)$ onto the manifold over 2D image pairs X and a function σ_z that maps z_i onto the manifold over 3D DVFs Z. Intuitively, these manifolds can be considered as atlases of respiratory motion. The central hypotheses of the present work are first, there exists a function f that maps from the manifold over 2D image pairs X to the manifold over 3D DVFs Z and, second, that the composite transformation $z_i = \sigma_z^{-1} \cdot f \cdot \sigma_x\left(x_i^\alpha, x_{ref}^\alpha\right)$ over the joint manifold $X \times Z$ can be learned by a deep neural network. We test these hypotheses by evaluating the performance of a trained network on imaging data from two lung cancer patients.

Results

A key motivator for the use of Voxelmap in image-guided radiotherapy is to understand how insights from the labor-intensive tasks of segmentation and treatment planning should be updated during treatment. With this in mind, Voxelmap trains in a patient-specific manner using only data acquired on the planning day (Fig 1d). To validate this method, a deep neural network was trained and tested on imaging data acquired on two separate days. During each forward-pass, the neural network first

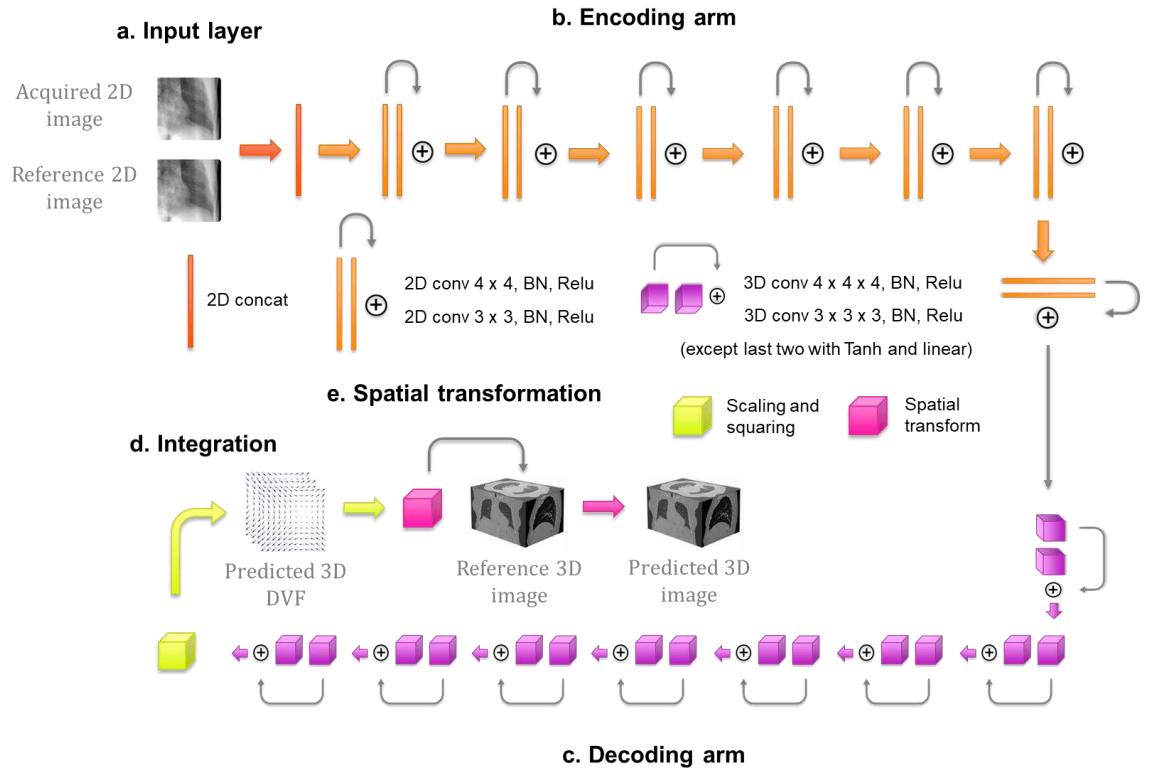

Fig 2. | Architecture of the Voxelmap network a, The input layer concatenates 2D image pairs **b**, The encoding arm consists of *n* repeating residual blocks (where 2D image size = $2^n \times 2^n$) followed by reshaping to a 3D tensor. **c**, The decoding arm consists of *n* repeating residual blocks followed by two additional convolutions at the desired 3D deformation vector field (DVF) size ($3 \times 2^n \times 2^n \times 2^n$) **d**, Optional integration layers can be included to encourage diffeomorphic transformations **e**, Predicted 3D images are produced by passing output 3D DVFs and reference 3D images through spatial transformation layers.

concatenates acquired and reference 2D images. This 2D image pair is then fed into an encoding arm, which extracts key features to create a latent low-dimensional representation of 2D image space. This low-dimensional feature map is then reshaped to a 3D tensor for processing by a decoding arm to produce a 3D DVF. Since respiratory motion should vary smoothly, we include the additional constraint that the underlying manifolds must be differentiable and therefore that the desired transformations be diffeomorphic. This constraint is imposed implicitly by using scaling and squaring layers[26] to efficiently integrate the output of the decoding arm. The resulting 3D DVF is passed through spatial transformation layers along with a reference 3D image to produce a predicted 3D image (Fig 2).

To simulate imaging conditions typically encountered in radiotherapy, Monte Carlo methods were used to create images with scatter and noise [27]. Network optimization then proceeded by taking these scatter- and noise-corrupted 2D images as input and minimising the mean-squared error (MSE) loss between predicted and ground-truth 3D DVFs. To encourage the network to focus on learning respiratory motion, this loss was computed within a thoracoabdominal mask. Additionally, 10 % of the training data was held-out during optimization for validation. Loss curves (Fig. 3) indicate that the models fit seen training data well and also generalized to unseen validation data.

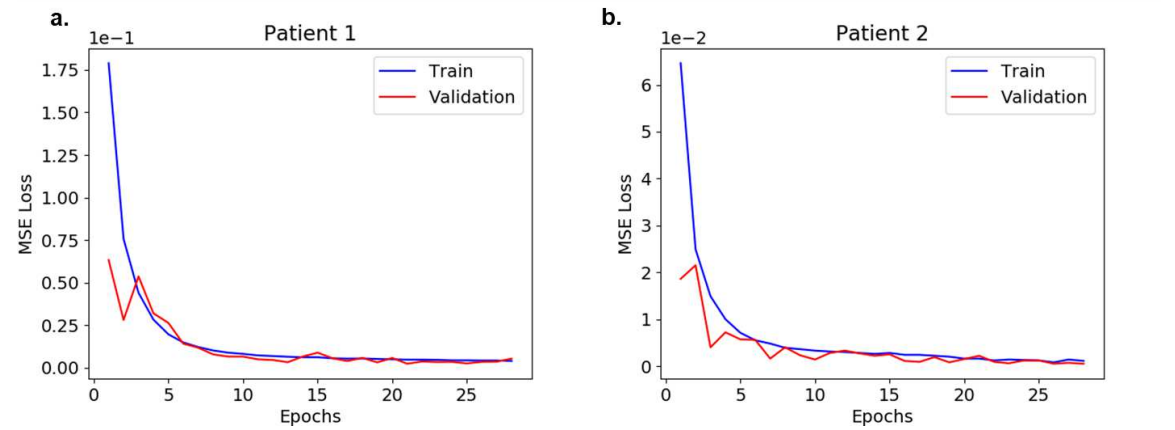

Fig 3. | Train and validation curves a,**b**, Plots of mean-squared error (MSE) loss for training data (in blue) and validation data (in red) against the number of epochs for Patient 1 (**a**) and Patient 2 (**b**).

Once trained, the neural network was the tested on imaging data corresponding to the treatment day. A clinician contoured a planning target volume (PTV) indicating the region to be treated for each patient. We evaluated the tumor tracking performance of Voxelmap by comparing the average DVF within this PTV. As shown in Figure 4, there was significant overlap between the predicted and ground-truth tumor positions. Patient 1 had mean tumor position errors of -0.08, -0.16, and -0.01 voxels along the left-right (LR), superior-inferior (SI) and anterior-posterior (AP) axes respectively (Table 1). Similarly, Patient 2 had errors of -0.13, -0.54, and 0.03 voxels respectively. Importantly, there were no noticeable differences in accuracy with changes in imaging angle. These results suggest that Voxelmap could be used to track respiratory-induced tumor motion during image-guided radiotherapy where the gantry continuously rotates around the patient as radiation is delivered.

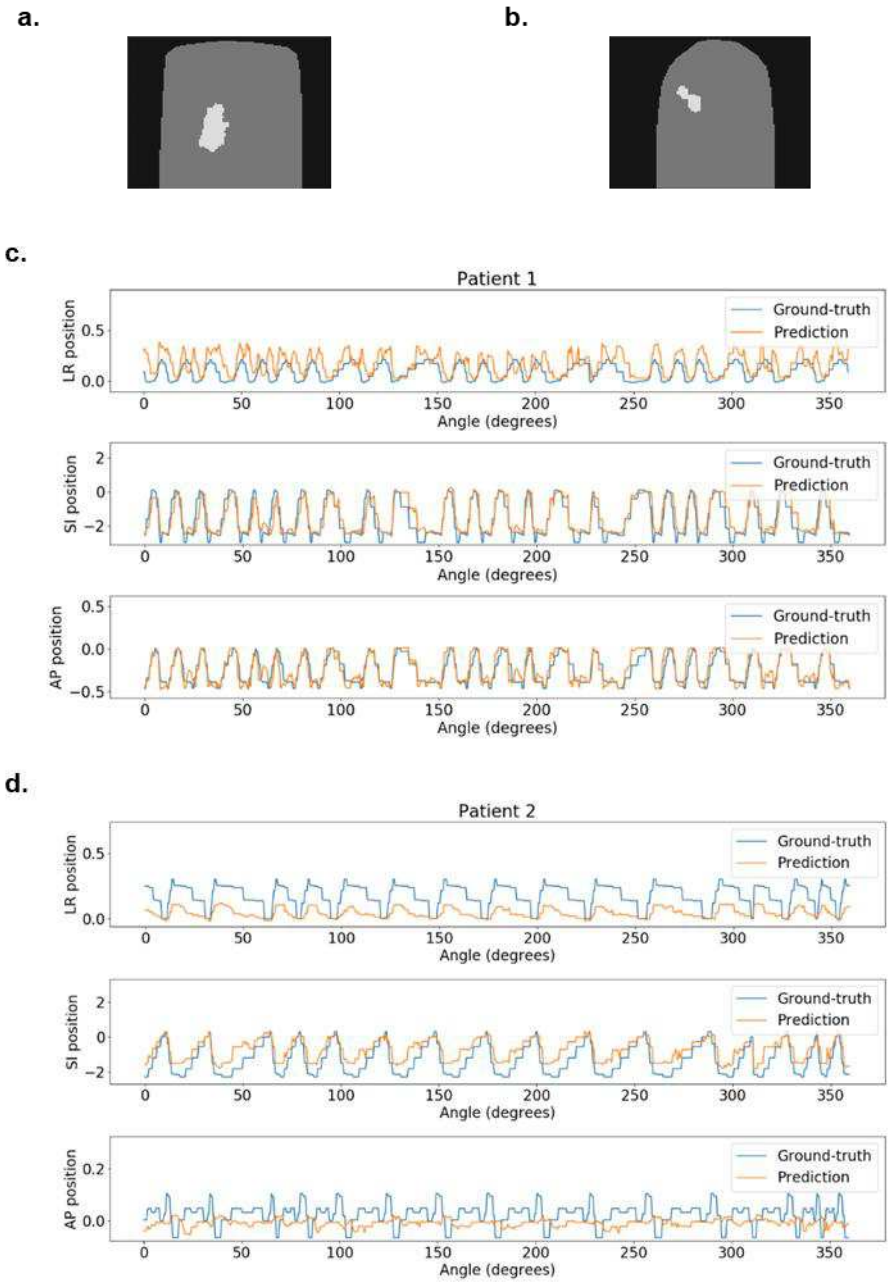

Fig 4. | Tumor tracking performance a,**b**, Thoracoabdominal (in dark gray) and planning target volume masks (in light gray) for Patient 1 (**a**) and Patient 2 (**b**) **c**,**d**, Plots of ground-truth (in blue) and predicted (in orange) tumor positions vs gantry angle along the left-right (LR), superior-inferior (SI) and anterior-posterior (AP) axes, where 0 corresponds to the position of the tumor on the reference 3D image for Patient 1 (**a**) and Patient 2 (**b**).

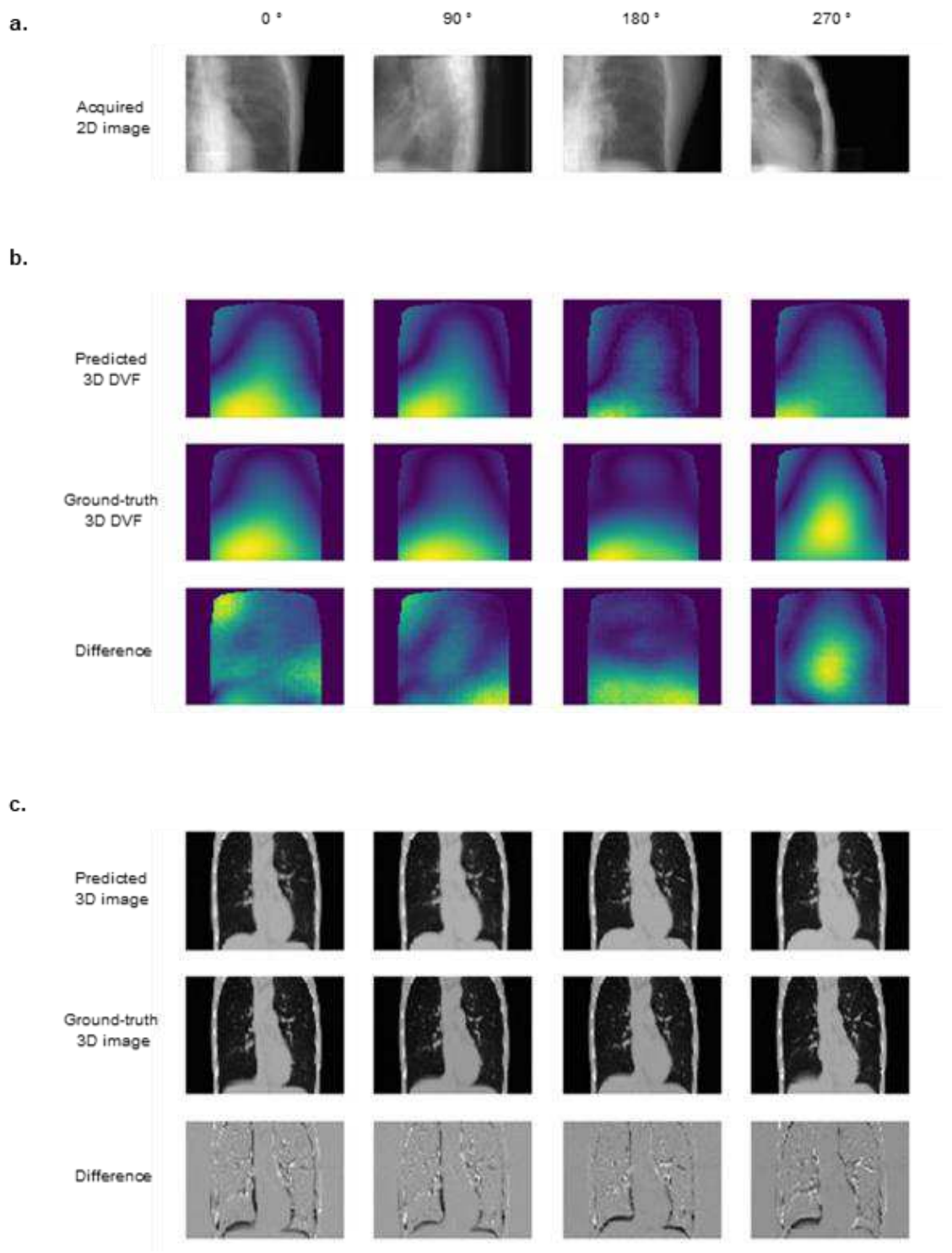

Fig 5. | Respiratory motion estimation and volumetric imaging for Patient 1 a, Acquired 2D images at 0°, 90°, 180° and 270° to showcase the performance of Voxelmap at different angles **b**, Predicted, ground-truth and difference 3D DVFs, where the magnitude of 3D motion is shown as heatmaps in the coronal axis **c**, Predicted, ground-truth and difference 3D images in the coronal axis.

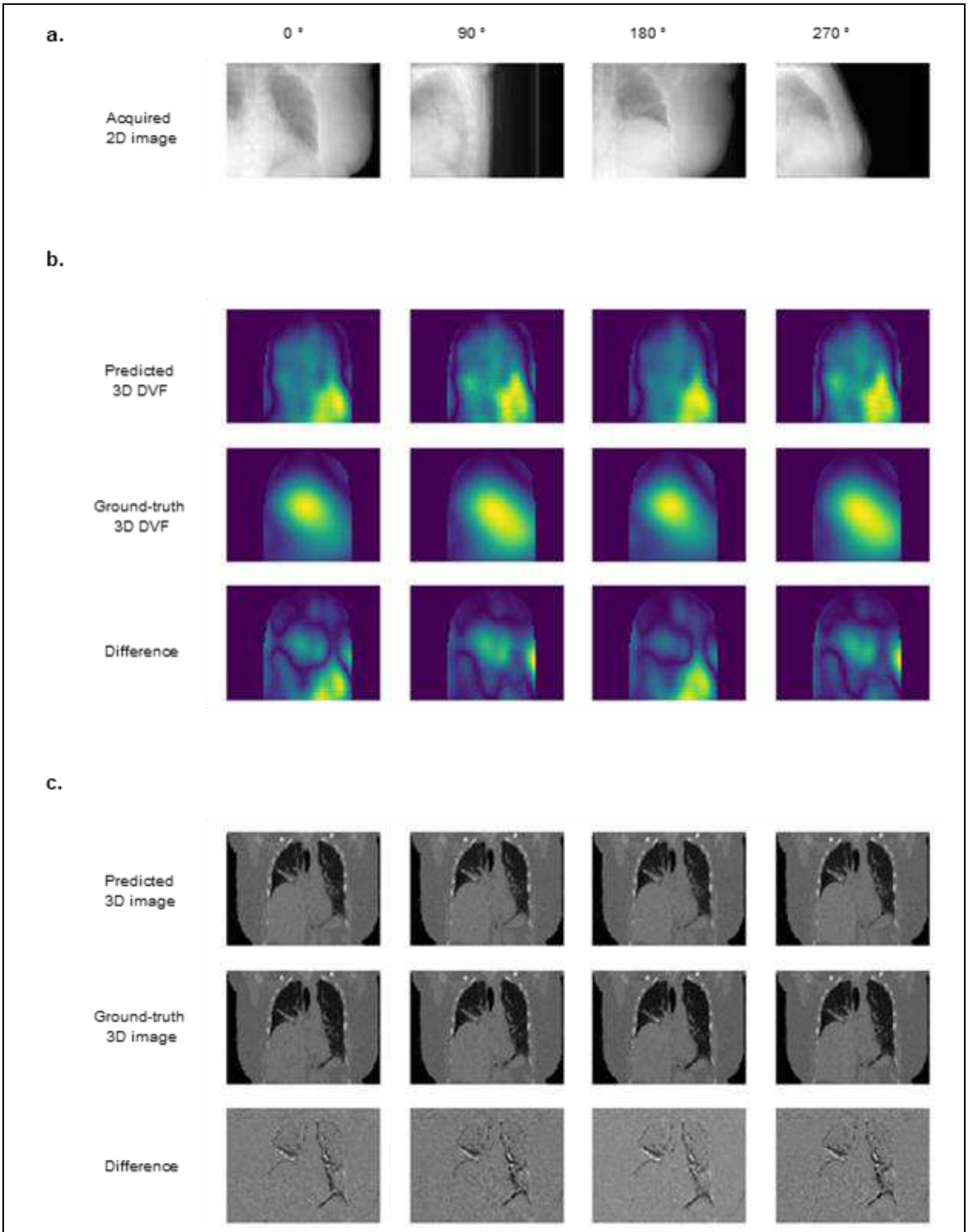

Fig 6. | Respiratory motion estimation and volumetric imaging for Patient 2 a, Acquired 2D images at 0°, 90°, 180° and 270° to showcase the performance of Voxelmap at different angles **b**, Predicted, ground-truth and difference 3D DVFs, where the magnitude of 3D motion is shown as heatmaps in the coronal axis **c**, Predicted, ground-truth and difference 3D images in the coronal axis.

Table 1 | Summary of performance on treatment day

	Image	Tumour motion			Thoracoabdominal motion		
Patient	**NRMSE**	**LR**	**SI**	**AP**	**LR**	**SI**	**AP**
1	0.09 %	-0.08	-0.16	-0.01	-0.11	0.01	-0.02
2	0.05 %	0.13	-0.54	0.03	0.06	-0.07	-0.07

NRMSE, normalized root-mean-squared error; LR, left-right error; SI, superior-inferior error; AP, anterior-posterior error. Image errors were measured in pixel intensity. Motion errors were measured in voxels.

Similarly, Voxelmap had mean errors in thoracoabdominal motion of less than one voxel for both patients. Comparing the predicted and ground-truth DVFs of Patient 1 at multiple angles, minor differences were observed except at 270° where the diaphragm and much of the thoracoabdominal anatomy was precluded from view (Fig 5). Conversely, for Patient 2, the field-of-view was such that the diaphragm was visible throughout the scan and differences between the predicted and ground-truth DVFs did not appear to depend on imaging angle (Fig 6). However, the DVFs produced by Voxelmap for Patient 2 appear to have more detail than those in the ground-truth. We suggest that this occurs because DVF smoothness is regularized in the conventional registration method[28] used to compute the ground-truth but such regularization was not imposed in the Voxelmap learning process. As a result, when the DVFs produced by Voxelmap were applied to reference images to generate 3D images much of the fine-structure was captured and there were only minor differences to the ground-truth. On the other hand, deviations from the ground-truth images were observed for Patient 1 near the diaphragm. Nonetheless, the normalised root-mean-squared error in normalised pixel intensities for these images was 0.09 % and 0.05 % for Patient 1 and Patient 2 respectively.

Discussion

To gain insight into the semantic representations learned by the Voxelmap network, activations from the first residual block were compared before and after training (Fig 7). These activation maps indicate that the network employs broad activation over the entire image prior to training, but focuses on specific anatomical regions of the once trained. In particular, the trained activation maps appear to favour high-contrast structures that are biomechanically important in thoracoabdominal motion, such as the ribs and intercostal muscles. Additionally, strong activations were observed in the lower-most pixels of the image over which the diaphragm deforms during respiration. This sparse representation of respiratory motion reflects the notion that 3D DVFs can be learned by mapping from a manifold over 2D image pairs. Indeed, these activation maps indicate that the trained network extracts key features in 2D image space to estimate respiratory motion.

Overfitting is a perennial challenge in machine learning that occurs when a large number of parameters are optimized to fit seen data but do not generalize well to unseen data. Voxelmap addresses this challenge by training neural networks in a context-specific manner. The core idea behind our proposed framework is that the problem of mapping from 2D images to 3D motion can be solved by learning manifold representations that reflect the specific biomechanics of the context-of-interest. In this paper, our method leverages the accuracy and specificity of optimizing over a large number of parameters for a particular patient while avoiding the issue of generalization by never using the same parameters across different patients. In image-guided radiotherapy, training data can be produced abundantly for this purpose by forward-projecting 3D images acquired during pre-treatment

scans. Here a 4D-CT was acquired for each patient yielding 10 3D images that were then each projected at 680 different angles, yielding almost 7000 training examples.

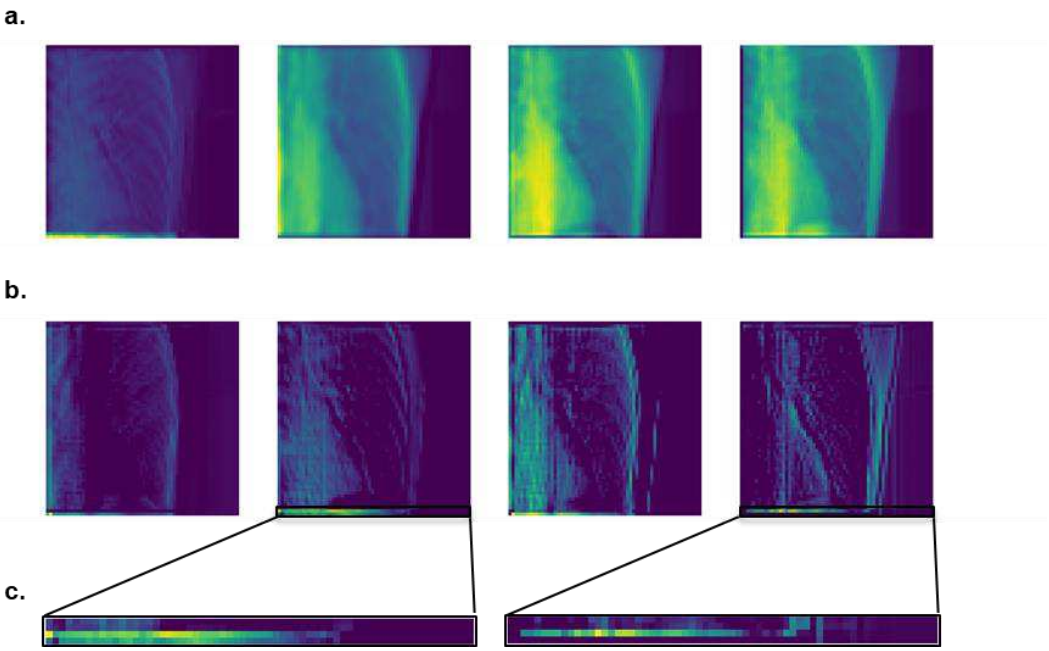

Fig 7. | Analysis of activations in the Voxelmap network a,**b**, Visualisation of the activation maps across the 4 output channels of the first residual block **(a)** before training and **(b)** after training with **c**, a zoom-in of the strong activations shown in the lower-most pixels of the trained maps. Brighter pixels indicate stronger activation.

One limitation of the present work is that Voxelmap was trained and tested on a unique dataset where the desired reference images are known on the day of treatment. However, the appropriate reference images can be produced using images available in existing clinical workflows. In particular, a pre-treatment scan of the patient is routinely acquired to determine whether any significant anatomical differences have occurred intervening time from the planning day. This scan can be used to deform a reference 3D image from the planning day to account for these anatomical variations and supply a source image on the day of treatment. Another limitation is that DVFs computed using conventional registration methods were used as ground-truth for respiratory motion. There is no way of knowing how every thoracoabdominal structure translates, rotates and deforms precisely given the corresponding image data. However, these conventional registration algorithms are used routinely to account for anatomical differences in the course of radiotherapy treatment. Voxelmap leverages the implicit constraints on patient-specific respiratory motion to make this information available and actionable during treatment.

Existing deep learning approaches to volumetric imaging map from 2D x-ray projections to 3D computed tomography images without predicting 3D motion[29,30]. However, motion is constrained in ways that images are not. One of the core motivations for the proposed framework is that biologically-

driven motion exhibits discoverable patterns. By mapping to a constrained solution space, Voxelmap was able to continuously estimate respiratory-induced tumor motion despite changing imaging angles. This flexibility is essential in the context of interventional and diagnostic procedures where images are acquired at many different angles. In contrast, previous volumetric imaging methods required the training of a new network on a different dataset for each imaging angle. In spite of its flexibility, Voxelmap employs a much smaller network than that of Shen *et al.*[29] for instance. Indeed, for the same image size, our proposed framework required the storage of 50-fold fewer trainable parameters ($1x10^7$ vs $5x10^8$), thereby drastically decreasing memory requirements. Our lightweight network also performed inference in 50 ms suggesting the possibility of real-time implementation. Additionally, existing volumetric imaging techniques were validated on "clean" digitally reconstructed radiographs, while the networks in this paper were trained and tested on scatter- and noise-corrupted images that reflect imaging conditions typically encountered in clinical scenarios.

Outlook. In this paper we introduce Voxelmap, a deep learning framework for real-time 2D-3D image registration and volumetric imaging in the presence of biologically-driven motion. Here we demonstrated the use of Voxelmap in the context of image-guided radiotherapy. A majority of radiotherapy centers around the world wish to implement real-time motion management during treatment but are hindered by finances, human resources and machine capacity. Voxelmap addresses this lacuna by using neural networks to learn patient-specific breathing patterns from the images available in existing clinical workflows. While the present work used Voxelmap for x-ray guided radiotherapy, this framework could be leveraged for other imaging modalities and applications. For instance, the proposed method of respiratory motion estimation could also be used in MRI-guided radiotherapy, which leverages the enhanced soft-tissue contrast of magnetic resonance imaging for tumor targeting. Moreover, the volumetric imaging capabilities of Voxelmap could enable clinical teams to visualize the changing 3D internal anatomy of their patients during interventional and diagnostic procedures for the first time.

Methods

Problem formulation. Mapping from 2D images to biologically-driven 3D motion is formulated as a problem of manifold approximation via deep learning. In particular, we hypothesize that the composite transformation $z_i = \sigma_z^{-1} \cdot f \cdot \sigma_x\left(x_i^\alpha, x_{ref}^\alpha\right)$ can be learned by a deep neural network where σ_x that maps every 2D image pair $\left(x_i^\alpha, x_{ref}^\alpha\right)$ onto the manifold X, σ_Z that maps every 3D DVF z_i onto the manifold Z and f that maps between X and Z. Here 2D image pair $\left(x_i^\alpha, x_{ref}^\alpha\right)$ is acquired by forward-projecting 3D image pair $\left(y_i, y_{ref}\right)$ at angle α and $\left(y_i, y_{ref}\right)$ are related by 3D DVF z_i. This hypothesis is tested by evaluating network performance on imaging data from two lung cancer patients.

Imaging data. Imaging data was acquired using scans from the Sparse-view reconstruction (SPARE) challenge[27]. Briefly, 1-min CBCT scans were simulated using 4D-CT volumes for patients with locally advanced non-small-cell lung cancer receiving 3D conformal radiotherapy. 4D-CT volumes used to simulate planning were acquired on a different day to those used to simulate treatment. For the planning day data, the 4D-CT supplied 10 3D images each which were projected to produce 2D images at 680 angles for one 360° revolution. This yielded 6800 volume-projection pairs with a 90:10

training:validation split - i.e. 6120 images for training and 680 images for validation. For the treatment day data, respiratory motion was simulated by converting real-time position management (Varian Medical Systems, Palo Alto, US) traces into respiratory phases and acquiring projections for the corresponding 4D-CT volumes at 680 angles for one 360° revolution. This yielded 680 volume-projection pairs for testing. All projections were simulated for a 120 kVp beam with a pulse length of 20 ms going through a half-fan bowtie filter. Scatter and noise were generated at 40 mA tube current via Monte Carlo methods to simulate imaging conditions commonly encountered in radiotherapy. Of the nine patients originally included in the SPARE study, Monte Carlo simulations with the same scatter and noise profiles were only generated for two patients over two separate imaging days. We validate our proposed method against the imaging data for both of these patients in this study.

Every 2D projected image was initially generated with pixel sizes of 0.776 mm × 0.776 mm and dimensions 512 × 384. However, they were downsampled to 128 × 128 due to memory constraints. Similarly, every 3D volumetric image was initially generated voxel sizes of 1 mm × 1 mm × 1 mm and dimensions 450 × 220 × 450 but was resized to 512 × 256 × 512 by cubic interpolation and downsampled to 128 × 128 × 128. Resizing in this manner means that motion of less than 3 mm will not be captured accurately, but this merely reflects current hardware limitations since the network can be easily scaled to accommodate larger image sizes as computational technology advances. Pixel intensities for all images were normalised between 0 and 1. Lung and planning target volume (PTV) masks were delineated by a clinician for each patient on the peak-exhalation 4D-CT. Lung masks were used to generate thoracoabdominal masks by generating a convex hull to include both lungs, expanding the resulting hull by binary dilation to include the ribs and extension inferiorly to the bottom of the image. Deformable image registration between the peak-exhalation 4D-CT and every other image of the 4D-CT for the simulated planning and treatment days was performed using the Elastix toolkit[28] to produce 3D DVFs for training and testing.

Neural network architecture. The neural network consists of 5 components: (1) an input layer (2) an encoding arm (3) a decoding arm (4) integration layers (5) a spatial transformation module (Figure 2). To describe the rationale behind each of these components: the input layer concatenates 2D image pairs; the encoding arm serves to produce a latent, low-dimensional representation of the key features in 2D image pairs; conversely, the decoding arm serves to map from this latent, low-dimensional representation to the 3D DVF between 3D image pairs; the integration layers encourage diffeomorphic transformations by computing the integral of the output from the final layer of the decoding arm; lastly, the spatial transformation module serves to deform the reference 3D image using the output 3D DVF of the neural network, enabling volumetric imaging.

In this study, 2D image pairs are concatenated to produce an input tensor of dimensions 2 × 128 × 128, where the first number indicates the number of channels while the second and third indicate image size. This is fed into the encoding arm of the neural network, which consists of n = 7 (since image size = 128 × 128 = $2^7 \times 2^7$) repeating residual blocks where each input tensor is convolved with a kernel of size 4 × 4 with stride 2 and padding 1, followed by a kernel of size 3 × 3 with stride 1 and padding 1 and batch normalization. This repeating pattern of convolutions yields output images at half the dimension of the original inputs and the number of output channels is chosen as double that of the original input. Hence, the first residual block produces an output tensor of dimension 4 × 64 × 64, the next residual block produces an output tensor of dimension 8 × 32 × 32 and so on until the final block of the encoding arm with output dimensions 256 × 1 × 1. Importantly, the number of residual blocks is determined intrinsically by the size of the input images such that larger images with detailed DVFs are processed by larger networks. The output tensor of the encoding arm is taken to represent a low-

dimensional latent representation of the 2D DVFs between projection pairs, which is reshaped to dimensions $256 \times 1 \times 1 \times 1$ for processing by the decoding arm.

In a reciprocal manner to that of the encoding arm, the decoding arm also consists of n = 7 repeating residual blocks. However, each block performs transpose convolution with a kernel of size $4 \times 4 \times 4$ with stride 2 and padding 1, followed by a kernel of size $3 \times 3 \times 3$ with stride 1 and padding 1 and batch normalization. Two additional convolutions are then performed at the desired output image size of $128 \times 128 \times 128$ with kernels of size $3 \times 3 \times 3$ with stride 1 and padding 1, and the number of channels is chosen to produce a final output tensor of dimensions $3 \times 128 \times 128 \times 128$. Every convolution, transpose convolution and fully connected layer uses ReLu activation, except the penultimate and final layers which use Tanh and linear activation respectively.

The output of the decoding arm is then fed through scaling and squaring layers[26] (step size = 10), which efficiently integrate the corresponding tensor. This process yields a 3D DVF, which is then fed into a spatial transform module along with the reference 3D image to produce a predicted 3D image.

Neural network training. Training loss was defined as $MSE(z_{true}, z_{pred})$, where MSE is mean squared error, z_{true} is the true 3D DVF and z_{pred} is the predicted 3D DVF. To encourage the network to focus on learning respiratory motion, this loss was computed within a thoracoabdominal mask. The neural network was trained using the Adam learning algorithm with learning rate 1×10^{-5} and batch size 4 for 30 epochs, requiring approximately 20 h on a NVIDIA RTX 2080 Super Max-Q 8 GB GPU.

Evaluation metrics. Once trained, the network was tested using simulated treatment images. Importantly, these images and the corresponding DVFs were unseen during training. To evaluate image accuracy, normalised root-mean-squared error in pixel intensities between the ground-truth and predicted 3D volumetric images were recorded. To evaluate DVF accuracy, mean errors in voxels between the ground-truth and predicted 3D DVFs were measured along the left-right, superior-inferior, and anterior-posterior axes within thoracoabdominal and PTV masks. Lastly, to evaluate network efficiency, mean inference time and the number of trainable parameters were recorded.

Acknowledgements

N.H. acknowledges funding from the Australian Government, the Australian-American Fulbright Commission and the Kinghorn Foundation. P.K. acknowledges funding from an Australian Government NHMRC Investigator Grant. The authors would also like to thank Dr Helen Ball, Dr Brendan Whelan, and Dr Paul Liu for their assistance in drafting the manuscript.

Author contributions

N.H. devised the deep learning framework, carried out all experimental work, and wrote the manuscript. C.C.S. supplied the validation data, wrote image processing code, assisted with experimental design, and reviewed the manuscript. P.K. conceived the applications to image-guided radiotherapy, assisted with experimental design, and reviewed the manuscript.

Competing interests

N.H., C.C.S. and P.K. are inventors on a filed provisional patent for the method disclosed in this paper.

Data availability

The datasets used in this work are publically available at image-x.sydney.edu.au/spare-challenge.

References

1. Huettel SA, Song AW, McCarthy G. *Functional magnetic resonance imaging.* Vol 1: Sinauer Associates Sunderland, MA; 2004.
2. Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann A. Neurophysiological investigation of the basis of the fMRI signal. *Nature.* 2001;412(6843):150-157.
3. Belliveau JW, Kennedy DN, McKinstry RC, et al. Functional Mapping of the Human Visual Cortex by Magnetic Resonance Imaging. *Science.* 1991;254(5032):716-719.
4. Rousseau F, Glenn OA, Iordanova B, et al. Registration-Based Approach for Reconstruction of High-Resolution In Utero Fetal MR Brain Images. *Academic Radiology.* 2006;13(9):1072-1081.
5. Lloyd DFA, Pushparajah K, Simpson JM, et al. Three-dimensional visualisation of the fetal heart using prenatal MRI with motion-corrected slice-volume registration: a prospective, single-centre cohort study. *The Lancet.* 2019;393(10181):1619-1627.
6. Verellen D, Ridder MD, Linthout N, Tournel K, Soete G, Storme G. Innovations in image-guided radiotherapy. *Nature Reviews Cancer.* 2007;7(12):949-960.
7. Murphy MJ, Balter J, Balter S, et al. The management of imaging dose during image-guided radiotherapy: Report of the AAPM Task Group 75. *Medical Physics.* 2007;34(10):4041-4063.
8. Goshtasby AA. *2-D and 3-D image registration: for medical, remote sensing, and industrial applications.* John Wiley & Sons; 2005.
9. Barton MB, Frommer M, Shafiq J. Role of radiotherapy in cancer control in low-income and middle-income countries. *The Lancet Oncology.* 2006;7(7):584-595.
10. Barton MB, Jacob S, Shafiq J, et al. Estimating the demand for radiotherapy from the evidence: A review of changes from 2003 to 2012. *Radiotherapy and Oncology.* 2014;112(1):140-144.
11. Tyldesley S, Delaney G, Foroudi F, Barbera L, Kerba M, Mackillop W. Estimating the Need for Radiotherapy for Patients With Prostate, Breast, and Lung Cancers: Verification of Model Estimates of Need With Radiotherapy Utilization Data From British Columbia. *International Journal of Radiation Oncology, Biology, Physics.* 2011;79(5):1507-1515.
12. Ferlay J EM, Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F. Global Cancer Observatory: Cancer Today. *Lyon, France: International Agency for Research on Cancer.* 2020.
13. Keall PJ, Mageras GS, Balter JM, et al. The management of respiratory motion in radiation oncology report of AAPM Task Group 76a). *Medical Physics.* 2006;33(10):3874-3900.
14. Kilby W, Dooley JR, Kuduvalli G, Sayeh S, Maurer CR. The CyberKnife® Robotic Radiosurgery System in 2010. *Technol Cancer Res Treat.* 2010;9(5):433-452.
15. Kamino Y, Takayama K, Kokubo M, et al. Development of a four-dimensional image-guided radiotherapy system with a gimbaled X-ray head. *International Journal of Radiation Oncology, Biology, Physics.* 2006;66(1):271-278.
16. Kupelian P, Willoughby T, Mahadevan A, et al. Multi-institutional clinical experience with the Calypso System in localization and continuous, real-time monitoring of the prostate gland during external radiotherapy. *International Journal of Radiation Oncology, Biology, Physics.* 2007;67(4):1088-1098.

17. Shah AP, Kupelian PA, Willoughby TR, Meeks SL. Expanding the use of real-time electromagnetic tracking in radiation oncology. *Journal of applied clinical medical physics.* 2011;12(4):3590-3590.
18. O'Shea T, Bamber J, Fontanarosa D, van der Meer S, Verhaegen F, Harris E. Review of ultrasound image guidance in external beam radiotherapy part II: intra-fraction motion management and novel applications. *Physics in Medicine and Biology.* 2016;61(8):R90-R137.
19. Lagendijk JJW, Raaymakers BW, van Vulpen M. The Magnetic Resonance Imaging–Linac System. *Seminars in Radiation Oncology.* 2014;24(3):207-209.
20. Mutic S, Dempsey JF. The ViewRay System: Magnetic Resonance–Guided and Controlled Radiotherapy. *Seminars in Radiation Oncology.* 2014;24(3):196-199.
21. Anastasi G, Bertholet J, Poulsen P, et al. Patterns of practice for adaptive and real-time radiation therapy (POP-ART RT) part I: Intra-fraction breathing motion management. *Radiotherapy and Oncology.* 2020;153:79-87.
22. Zhu B, Liu JZ, Cauley SF, Rosen BR, Rosen MS. Image reconstruction by domain-transform manifold learning. *Nature.* 2018;555(7697):487-492.
23. Ford EC, Mageras GS, Yorke E, Ling CC. Respiration-correlated spiral CT: A method of measuring respiratory-induced anatomic motion for radiation treatment planning. *Medical Physics.* 2003;30(1):88-97.
24. Low DA, Nystrom M, Kalinin E, et al. A method for the reconstruction of four-dimensional synchronized CT scans acquired during free breathing. *Medical Physics.* 2003;30(6):1254-1263.
25. Vedam SS, Keall PJ, Kini VR, Mostafavi H, Shukla HP, Mohan R. Acquiring a four-dimensional computed tomography dataset using an external respiratory signal. *Physics in Medicine and Biology.* 2002;48(1):45-62.
26. Dalca AV, Balakrishnan G, Guttag J, Sabuncu MR. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical Image Analysis.* 2019;57:226-236.
27. Shieh C-C, Gonzalez Y, Li B, et al. SPARE: Sparse-view reconstruction challenge for 4D cone-beam CT from a 1-min scan. *Medical Physics.* 2019;46(9):3799-3811.
28. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. elastix: A Toolbox for Intensity-Based Medical Image Registration. *IEEE Transactions on Medical Imaging.* 2010;29(1):196-205.
29. Shen L, Zhao W, Xing L. Patient-specific reconstruction of volumetric computed tomography images from a single projection view via deep learning. *Nature Biomedical Engineering.* 2019;3(11):880-888.
30. Lei Y, Tian Z, Wang T, et al. Deep learning-based real-time volumetric imaging for lung stereotactic body radiation therapy: a proof of concept study. *Physics in Medicine & Biology.* 2020;65(23):235003.