Beyond respiratory models: a physics-enhanced synthetic data generation method for 2D-3D deformable registration

François Lecomte [1,†] Pablo Alvarez [1] Stéphane Cotin [1] Jean-Louis Dillenseger[2]
[1] INRIA, [2] University of Rennes
[1] {francois.lecomte, pablo.alvarez, stephane.cotin}@inria.fr, [2] jean-louis.dillenseger@univ-rennes.fr

Abstract

Deformable image registration is crucial in aligning medical images for various clinical applications, yet enhancing its efficiency and robustness remains a challenge. Deep Learning methods have shown very promising results for addressing the registration process, however, acquiring sufficient and diverse data for training remains a hurdle. Synthetic data generation strategies have emerged as a solution, yet existing methods often lack versatility and often do not represent well certain types of deformation. This work focuses on X-ray to CT 2D-3D deformable image registration for abdominal interventions, where tissue deformation can arise from multiple sources. Due to the scarcity of real-world data for this task, synthetic data generation is unavoidable. Unlike previous approaches relying on statistical models extracted from 4DCT images, our method leverages a single 3D CT image and physically corrected randomized Displacement Vector Fields (DVF) to enable 2D-3D registration for a variety of clinical scenarios. We believe that our approach represents a significant step towards overcoming data scarcity challenges and enhancing the effectiveness of DL-based DIR in a variety of clinical settings.

1. Introduction

Deformable image registration (DIR) refers to the process of finding a transformation between two (or more) medical images as to optimally align their underlying anatomical structures. This process has been a longstanding area of research in the medical imaging community, giving its great potential for a variety of clinical applications [12, 15]. Although many methods for DIR exist to date, researchers continue to investigate various ways to improve their efficiency and robustness, which currently hinder the widespread use of DIR in clinical settings. The advent of Deep Learning (DL) methods has enabled important improvements to the efficiency of DIR algorithms, since transformations between image pairs can be predicted very fast with acceptable accuracy [15].

However, a major difficulty of DL methods stems from the quality and amount of data needed for their training. In the context of DIR, these data typically take the form of high quality paired undeformed-deformed images for unsupervised approaches [6], ground-truth deformation fields for supervised approaches [9, 11], or a less restrictive combination of the two for weakly-supervised approaches [4]. Regardless of the chosen approach, it is often difficult to have access to sufficiently large experimental datasets, since they are nearly impossible to acquire in practice. This is particularly true in the context of intra-patient 2D-3D registration involving fluoroscopic and CT imaging, for which acquiring a sufficient number of image pairs would lead to very high X-ray radiation exposure. To cope with this problem, researchers typically adopt synthetic data generation strategies, which allow the constitution of large-enough datasets from a low number of medical images.

For instance, ground truth deformations for supervised training can be computed by interpolating between deformation modes in time varying 4DCT images [10], which are common in radiotherapy as it involves respiratory motion. However, clinical applications where such rich time varying data is unavailable cannot exploit this data generation strategy. Another approach to synthetic data generation consists in randomly generating deformation fields, which can then directly serve as ground truth for supervised learning [11], or as a way to produce deformed images for unsupervised learning [5]. However, it is crucial for this strategy to account for the whole range of possible deformations as to effectively train the DL algorithm.

In this work, we address the problem of X-ray to CT 2D-3D DIR for the assistance to fluoroscopy-guided abdominal interventions in the presence of large arbitrary deformation. We are thus interested in a clinical application where tissue deformation is not restricted to respiratory motion, but other sources of deformation such as the insertion of instruments (*e.g.* needle, catheter) and changes in patient position are

† Corresponding author.

possible. It is noteworthy to mention that no dataset exists for such clinical context (nor can it reasonably be acquired), and a synthetic generation strategy for training a DL method is therefore necessary. We further restrict ourselves to a single X-ray acquisition during the intervention, since multi-plane X-ray acquisition require specialized equipment that may not be available in all interventional suites.

To our best knowledge, few works have addressed the single X-ray to CT registration problem using DL methods. Foote et al. showed that a neural network could accurately predict deformation modes of a respiratory displacement dataset from Digitally Reconstructed Radiograph (DRR) images [3]. Also, Nakao et al. proposed a DL framework to predict abdominal organ shapes from DRRs, with an accuracy ranging from 3.5 mm to 6.1 mm at the organ surface [8]. However, this method only predicts displacements at the surface of organs, and displacements inside the organs remain unknown. Shao et al. proposed a DL framework in combination with a biomechanical model correction step for liver tumor localization, and reported errors ranging from 2.83 mm to 2.95 mm [10]. Nonetheless, all these mentioned DL methods were trained with synthetic data from statistical models built upon 4DCT respiratory images. The underlying assumption of this kind of generative approach is that the statistical model accurately represents the wide range of anatomical deformation possible during the intervention. This assumption may however be incorrect in clinical settings where surgically induced deformations are present, leading to degraded registration accuracy.

In this work, we propose a generic approach to synthetic data generation for training a DL 2D-3D DIR framework. As opposed to state-of-the-art methods in the literature, our method only requires a single 3D CT image for the generation of training data, and is agnostic to the registration problem at hand thanks to the use of physically corrected randomized Displacement Volume Fields (DVF). As a result, our method is equipped to address a variety of applications where large, nonrigid deformations may occur.

2. Method

2.1. Overview

Our data generation process is centered around commonly performed fluoroscopy-guided interventions, where a CT scan of the patient is acquired before the intervention, and used to plan the intervention.

First, a pre-operative CT scan is acquired and structures of interest are segmented. The intervention can then be planned by the clinicians. We assume that the pose of the C-arm with respect to the patient is determined during this step. Using only the pre-operative CT, structures of interest and the C-arm pose, a synthetic, domain-agnostic dataset is generated to train the neural network to recover an arbitrary deformation from a fluoroscopic image. Then, during the intervention, the C-arm is positioned as per planning and fluoroscopic images are acquired for intra-operative guidance. Each intra-operative image can then be augmented in real-time by updating the pre-operative data using the network, and projecting it on top of the image. This workflow is summarized in Fig. 1.

To train a neural network to estimate a deformation from a fluoroscopy, we use the pre-operative CT-Scan I and the C-arm pose P to generate a training dataset. This synthetic dataset is composed of pairs of synthetic deformations ϕ_i and DRR projections p_i. To generate each sample of the dataset, the process goes as follows: ϕ_i is generated using a sum of randomized Gaussian kernels (Sec. 2.2) and further processed to ensure a realistic range of deformations (Sec. 2.3). Then, ϕ_i is used to generate a deformed CT image I'_i from I (Sec. 2.4) and the corresponding synthetic fluoroscopic image p_i is generated from I'_i and P (Sec. 2.5). Finally, the neural network is then trained on the synthetic dataset (see Sec. 2.6).

In a clinical setting, at test time, a single fluoroscopic image of the patient would be acquired and processed in real-time by the network to register the pre-operative CT scan to the intra-operative anatomy. Thanks to the registration, structures of interest from the updated CT-Scan could be projected on the fluoroscopic image in real-time, enabling augmented intra-operative image guidance.

In our experiments, we instead used a post-operative CT-Scan of a porcine subject to generate a DRR displaying a realistic, surgery induced, deformation and used it to evaluate registration accuracy. We also evaluated the registration accuracy on synthetic CT scans where the ground truth deformation is known with perfect precision.

2.2. Deformation generation

A non-rigid deformation is defined on the 3D image $I(x)$ by $\phi(x) : \mathbb{R}^3 \to \mathbb{R}^3$ with x a point in the image volume and $\phi(x) = x + \varphi(x)$ a deformation, with $\varphi(x)$ a displacement vector field. We restrict the region where ϕ is defined to a sub-region of the volume, which will be referred to as the field domain. The key characteristics we seek in the displacement field are smoothness and invertibility. A good candidate for producing such displacement fields is the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework ([13]), which demonstrated very good performance in non-rigid registration problems ([2]). In this framework, the non-rigid deformation ϕ that registers an image I to an image I' is obtained by integrating a velocity field $V(t, x)$ over time, following a set of differential equations to drive the evolution of $V(t, x)$.

The authors demonstrate that $V(t, x)$ can be expressed

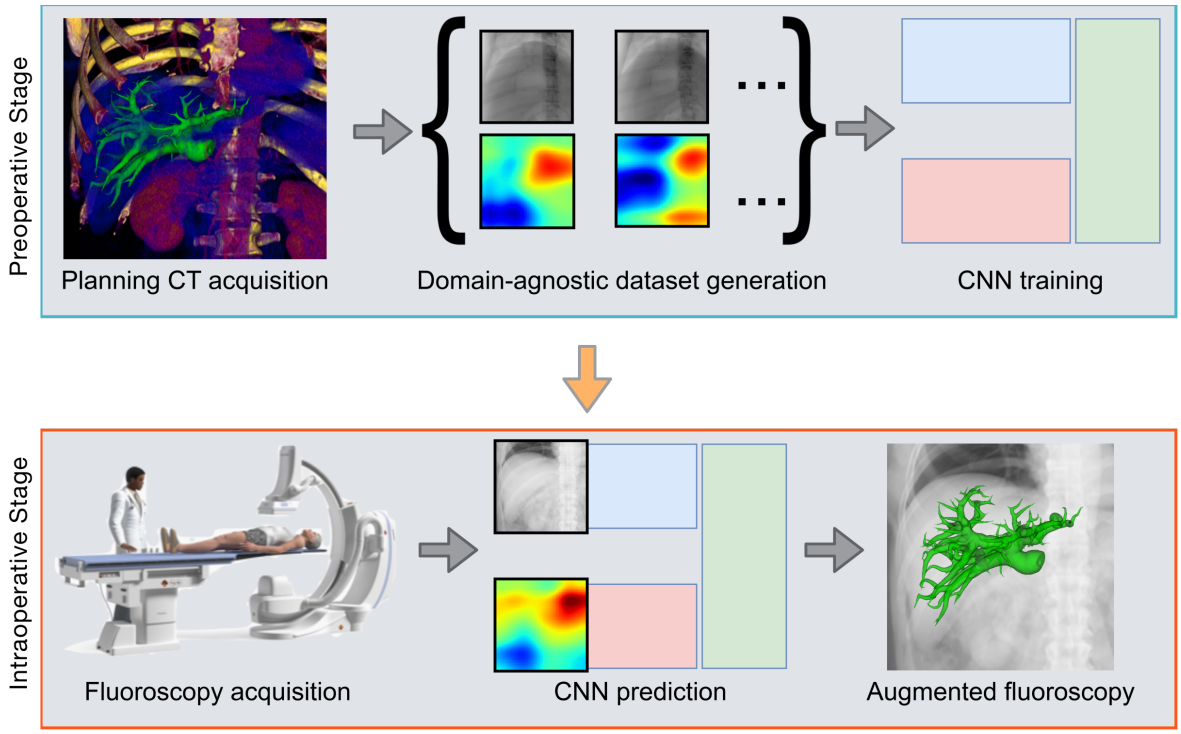

Figure 1. Overview of the proposed method. First, the intervention is planned from a 3D CT scan of the patient, where structures of interest are segmented and the C-arm pose is determined. Second, the neural network, detailed in Fig. 2, is trained on non-rigid deformations of the CT-Scan and synthetic fluoroscopic images. Third, during the intervention, the C-arm is positioned, and a fluoroscopic image is acquired. Fourth, the network computes the deformation from the fluoroscopic image and the warped segmentation is used to augment the fluoroscopy.

as:

$$V(t,x) = \sum_{k=1}^{N_{cp}} \alpha_k(t) \cdot K_k(x, y_k(t)) \tag{1}$$

where $K_k(t)$ are elements of a Reproducing Kernel Hilbert Space, such as Gaussian kernels, located at the N_{cp} control points $y_k \in \mathbb{R}^3$ and associated with weights $\alpha_k \in \mathbb{R}^3$. φ is then given by $\varphi(x) = \int_0^1 V(t,x)dt$.

In our framework, we directly compute $\varphi(x)$ by randomizing the control points y_k, covariance matrices $\sigma_k \in \mathbb{R}^{3\times3}$ and weights α_k of the Gaussian kernels.

$$\varphi(x) = \sum_{k=1}^{N_{cp}} \alpha_k \cdot K(x, y_k, \sigma_k) \tag{2}$$

To sample y_k, we generate a set of random points in the field domain. We then reject points that are closer than a threshold Δ_y in order to avoid sharp variations of φ and re-generate rejected points until the desired number of control points is obtained. α_k are sampled from a 3D uniform distribution. Finally, σ_k is generated as $N_y \times 3 \times 3$ i.i.d variables with values between 15% and 30% of the size of the field domain. To ensure that ϕ remains diffeomorphic, we verify that the spatial Jacobian $J(x) = \det(\nabla\phi(x))$ is bounded and positive accross the domain.

2.3. Deformation post-processing

While the randomly generated DVFs are smooth and diffeomorphic, they may still incompletely represent the range of possible deformation during the fluoroscopy-guided intervention, which is the consequence of two main factors.

First, the parameters for the Gaussian kernels are sampled independently for each kernel, meaning that in any given DVF, there will be both large and small deformations. This is potentially different from real deformations, which may in some cases be small throughout the domain. Obtaining such a small deformation DVF from our randomized generation process is very unlikely, since it would require all realizations for α_k to produce small values. To remediate this, the generated DVF is multiplied by a scaling factor between -1 and 1, which ensures that samples with overall small displacements are better represented in the dataset.

Second, since the DVF generation process is stochastic, there is no guarantee that a body deforming under the influence of such DVF respects the conservation laws of physics. We therefore correct the DVF with a biomechanical model.

We are only interested in the deformation of the liver's internal structures (*e.g.* tumor, vessels), and therefore correct the DVF only inside the region occupied by the liver, hereafter denoted by $\mathbf{\Omega}$. To that end, in a preprocessing step, we first perform the liver segmentation and meshing to obtain a tetrahedral mesh representing the liver domain $\mathbf{\Omega}$. Then, the displacement inside the liver $\mathbf{U}$ is computed as the solution to the nonlinear elastostatic problem:

$$-2\nabla \cdot \frac{\partial \Psi}{\partial \mathbf{C}} = \mathbf{0}, \text{ in } \mathbf{\Omega} \tag{3}$$

where $\mathbf{C} = \mathbf{F}^T\mathbf{F}$ is the right Cauchy-Green deformation tensor, Ψ is the strain energy density function, and the gradient of deformation tensor $\mathbf{F}$ is related to the displacement field $\mathbf{U}$ via $\mathbf{F} = \nabla\mathbf{U} + \mathrm{I}$.

The liver is modeled as a hyperelastic Neohookean solid with strain energy density function:

$$\Psi = \frac{\lambda}{4}(J^2 - 1 - 2\ln(J)) + \frac{\mu}{2}(I_C - 3 - 2\ln(J)), \tag{4}$$

where $J = \det(\mathbf{F})$, $I_C = \mathrm{tr}(\mathbf{C})$ is the first invariant of the right Cauchy-Green deformation tensor, and μ and λ are the so-called *Lamé parameters*.

The Finite Element Method (FEM) was used to solve the elastostatic problem (3), with Dirichlet boundary conditions extracted from the DVFs prescribed at the liver boundary. The corrected DVF is then obtained by composing the physically accurate displacement solution $\mathbf{U}$ from (3) inside the liver, and the DVF outside the liver. Since all the liver's boundary is constrained, we used $\mu = 1$ and $\lambda = 0$ for all our biomechanical simulations.

2.4. Image warping

We model the non-rigid deformations of the anatomy as coordinate transforms $\phi(x) = x + \varphi(x)$ with x a point in the CT image and φ a physically regularized Displacement Vector Field (DVF). The warped CT image $I'(x) = I \circ \phi(x)$ is obtained by linearly interpolating the values of I at $x' = \phi(x)$:

$$I'(x) = \sum_{z \in \mathcal{Z}(x')} I'(z) \prod_{d \in \{0,1,2\}} (1 - |x'_d - z_d|) \tag{5}$$

where z are the 8 voxels nearest to x' and d iterates through the 3 spatial components of x' and z. This operation is known as backward warping.

2.5. Digitally Reconstructed Radiographs

Once we have obtained the warped image I', we proceed by generating Digitally Reconstructed Radiographs (DRR) using the DeepDRR framework ([14]).

This framework models the C-arm as a pinhole camera, parameterized by a projection matrix P composed of an intrinsic matrix $H \in \mathbb{R}^{3\times3}$ and an extrinsic matrix $E \in \mathbb{R}^{3\times4}$. E is obtained from the planned pose of the C-arm and H is obtained from the characteristics of the C-arm detector panel. The fluoroscopic image p observed during the intervention is then approximated by:

$$p(u) \approx \int I'(x)\mathbf{dl_u} \tag{6}$$

with $\mathbf{l_u}(x) = P \cdot x$ the ray originating from the point $u \in \mathbb{R}^2$ on the detector plane. The DeepDRR framework uses a ray tracing algorithm that computes the line integral $p(u) = \int I'(x)\mathbf{dl_u}$ through $I'(x)$ for each pixel u of the 2D projection image $p(u)$, with l_u the 3D ray connecting the pixel u to the emission source. This results in the invariance of $p(u)$ to the distribution of $I(x)$ along the path of the ray. This is why displacements collinear to the projection rays cannot be directly observed in the projection image. In our case, the direction of the camera is aligned with the Antero-Posterior (AP) direction in the CT scan. Consequently, displacements in the AP direction are almost collinear with the projection rays, and will thus be almost invisible in the DRR.

2.6. Network architecture

The goal of the neural network is to register information from the pre-operative 3D CT Scan on the intra-operative 2D X-Ray image. Because of breathing and surgically-induced motion of the organs, the information from the pre-operative CT is outdated and needs to be updated before being projected on the intra-operative image. Consequently, in order to augment the intra-operative image, the network must learn to update the position of structures in the pre-operative CT using the intra-operative data. Practically, a X-Ray image is input to the network, which predicts the 3D displacement field that registers the CT to the X-Ray image.

Our fully convolutional network architecture (detailed in Fig. 2) is based on the architecture described in [7], with some modifications. With this architecture, 2D features are first extracted from the input 2D fluoroscopic image by a ResNet encoder, transformed into 3D features and decoded by a 3D convolutional decoder to obtain a displacement field in the 3D CT image space. The key characteristic of this architecture is the direct conversion from 2D to 3D feature maps. We improve on this architecture by using a back-projection module to perform the 2D to 3D conversion, instead of simply adding a spatial dimension by reshaping. This is necessary because the 2D image is acquired by an X-Ray detector, analogous to a pinhole camera. Thus, the position of objects in the 2D C-arm image is related to their position in the 3D CT image space via the camera matrix P. Owing to the local nature of convolutions, the feature maps extracted by the encoder are in the same space as the input image. So, in order to compute a displacement field in 3D CT image space from a 2D projective image, it is beneficial

to use the camera matrix to project the 2D feature maps to 3D space, thus preserving the spatial relationships between objects in the 2D to 3D conversion process.

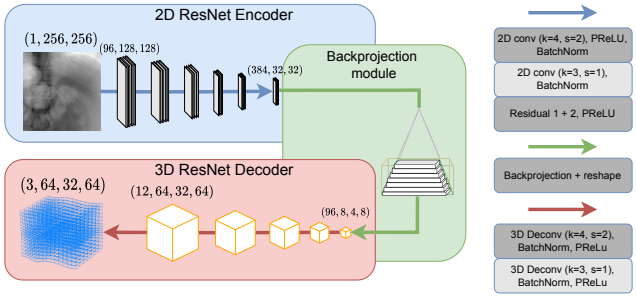

Figure 2. Each block in the Encoder downscales the feature maps and increases their number by a factor of 2. The backprojection module transforms the 2D feature maps into 3D feature maps. Every two layers in the Decoder, the number feature maps is divided by 2 and the spatial size is upsampled by a factor of 2. The last decoder layer transforms the 12 feature maps into a 3-channel 3D image, representing the 3D DVF.

Since the neural network receives a single 2D image as input, the 3D motion estimation task is inherently an ill-posed problem. In particular, the 3D motion component in the direction of the projection rays cannot be observed in the image. According to (6), any motion along the projection rays $\mathbf{l_u}$ will not induce a change in image intensity. Indeed, the value of the line integral along $\mathbf{l_u}$ is not modified by a voxel displacement along p_u.

Thus, we devise a projective loss $\mathcal{L}_{PMSE}$:

$$\mathcal{L}_{PMSE} = \left\| \mathcal{P}(\phi(x_i)) - \mathcal{P}(\hat{\phi}(x_i)) \right\|_2^2 \qquad (7)$$

where $\mathcal{P}$ is the projection operator that projects points in 3D space to points on the image plane using the camera matrix P, x_i are the 3D points on which the network is supervised (herein a regular grid of points around the structure of interest, with the points not visible in the projection masked in the loss computation), ϕ the displacement predicted by the network and $\hat{\phi}$ the ground truth displacement.

While the network is still limited in its ability to predict out-of-plane motion, using $\mathcal{L}_{PMSE}$ instead of a 3D MSE loss on the displacement field facilitates the learning of the network, because its prediction is supervised in the same space as its input, the 2D image space. In all experiments, the network was trained for 30 epochs, with the learning rate set at 5.10^{-5}, which took approximately 2 hours on an Nvidia RTX 4090 GPU.

3. Results

In order to validate our data generation approach, we evaluated the performances of a neural network trained on synthetic data, for two different registration contexts.

The first context was extracted from an open-source swine liver deformation dataset, IHUDeLiver10[1]. IHUDeLiver10 is composed of ten pairs of {*baseline; deformed*} Contrast Enhanced CT scans (CECT), experimentally acquired on ten different porcine subjects. For both images in each pair of CECT in the dataset, the portal vessel trees were segmented by an expert clinician, and serve to evaluate the registration accuracy. For each subject, the deformation of the anatomy was the result of a surgical procedure. Thus, this dataset contains realistic surgically-induced deformations of the anatomy, which can be used to validate the effectiveness of our synthetic data generation approach. To transform the Contrast Enhanced CTs into regular, non-contrasted CTs, we used image inpainting [1] to remove as much of the contrast effect due to contrast agents as possible. The pre-operative CT, *baseline* CT, was used to generate the training dataset, while the post-operative CT, *deformed* CT, was used to generate a test sample to evaluate the registration performance of the network. In this work, we only used one pair of experimentally acquired CT-Scans from the IHUDeLiver10 dataset (sample number 8). The deformation in the *deformed* CT of sample number 8 was induced by a surgical manipulation of the anatomy, reproducing deformations that may arise in a surgical intervention.

The second context was generated synthetically, in order to test the accuracy of the method in a more controlled setting, less dependent on anatomical particularities that may influence the performance of the network. The *baseline* synthetic CT is composed of a cube of the same volume and at the same position as the liver of the first test case, surrounded by voxels with a constant intensity corresponding that of skin tissue. Inside the cube, the voxel intensities alternate along a checkerboard pattern, with tiles of side length 13.75 mm. The intensity values remain constant within each tile, but they gradually increase across tiles along the cube's main diagonal. For this case, the test samples were generated by setting constant Dirichlet boundary conditions on the left and right faces of the cube (while leaving the remaining faces stress-free), and solving the elastostatic problem (3) using the FEM with an hexahedral mesh of side length 10 mm. The displacement on the left face of the cube was set to 0, while the displacement on the right face of the cube was set to -40 mm, -20 mm, +20 m and +40 mm along the Left-Right (LR) axis, respectively. To generate the test samples, the cube was modeled as a hyperelastic Mooney-Rivlin solid, instead of the simpler Neo-Hookean solid used to generate the training data, in order to avoid bias regarding the choice of the hyperelastic model in the test data. The deformed mesh was then used to interpolate displacements on the CT image and produce the

[1]IHUdeLiver10, along with data processing code, will be released at https://doi.org/10.57745/EUBXGH

deformed CT scans. A DRR was then generated for each deformed CT scan, as described before.

For both registration contexts, the C-arm pose P was defined such that the projection is centered on the liver, and the viewing direction of the C-arm was aligned with the Antero-Posterior (AP) anatomical axis. In the following experiments, each dataset contains 18000 training samples and 2000 validation samples. Since the proposed use of the method is augmented anatomical visualization on 2D fluoroscopic images, all errors were measured on the 2D image plane after projection with the operator $\mathcal{P}$.

For the liver registration context, no point-to-point correspondences were available between the *baseline* and the *deformed* vessel trees, and we therefore chose the Earth mover's distance (EMD) metric to evaluate the registration accuracy. For the second registration context, the points of the cube mesh were used to measure registration accuracy directly. Since this test case is generated synthetically and the points are paired between the *baseline* and *deformed* images, we used the reprojection distance metric (RPD) which measures the euclidean distance after projection (using $\mathcal{P}$) on the image plane.

The Figs. 3 and 4 show the baseline and deformed DRRs for the liver and synthetic contexts, respectively.

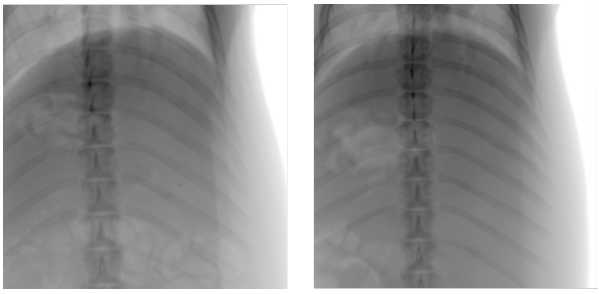

Figure 3. On the left, the DRR associated with the baseline CT and on the right the DRR associated with the deformed CT.

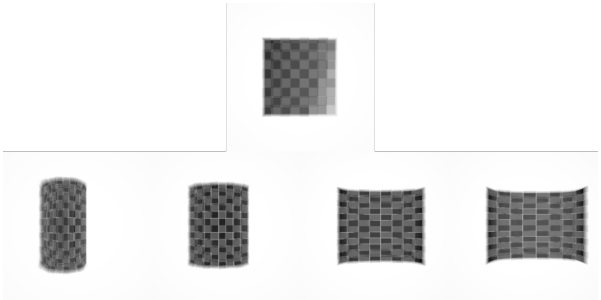

Figure 4. On top, the DRR associated with the baseline CT and on the bottom the DRRs associated with the deformed CTs, for displacements of -40 mm, -20 mm, +20 m and +40 mm (from left to right).

Epoch	Phy	Nophy
3	**4.0**	4.3
6	5.5	**3.6**
9	6.8	**4.7**
12	4.0	**3.9**
15	**4.2**	4.7
18	**3.4**	3.7
21	4.3	**3.3**
24	**2.8**	3.7
27	**3.7**	5.2
30	**3.6**	3.9

Table 1. Registration accuracy on the test sample every 3 epochs for networks trained with physically regularized (Phy) and not physically regularized (Nophy) data generation for the IHUDeLiver10 test case.

3.1. Registration accuracy

We evaluated the registration accuracy of the network trained using the synthetic data generation process described above.

The network was trained from scratch for each of the two test cases described above. For each test case, two training datasets were generated following 2.2, with and without the physical regularization step described in 2.3, in order to evaluate the effect of physics-based regularization.

For each test case, the registration accuracy of the network on the test sample(s) was measured every 3 training epoch. The tables 1 and 2 report the registration accuracy of the networks on the IHUDeLiver10 test sample and synthetic cube test samples respectively.

3.2. Ablation study

We performed two experiments on the IHUDeLiver10 test case to evaluate the impact of the data generation post-processing on the network performances. The architecture and training procedure of the network is the same for each experiment.

In the first experiment, three datasets were generated. The first dataset, termed "Base", was generated using the data generation process described above but without the post-processing described in Sec. 2.3. The second dataset, termed "Base + scale" was generated in the same way, but with the scaling post-processing and without the physical regularization. Finally, the third dataset, termed "Base + scale + phy" was generated using the full data generation process, with scaling and physical regularization, as described in Sec. 2.3.

The best registration performance for each dataset was 3.8 mm for the "Base" dataset, 3.3 mm for the "Base + scaling" dataset and 2.8 mm for the "Base + scaling + phy" dataset. In Fig. 5, the registration error of the network on

Stretching amount (mm)	-40		-20		20		40	
Epoch	Phy	Nophy	Phy	Nophy	Phy	Nophy	Phy	Nophy
3	**17.79**	19.83	**5.81**	7.74	**5.10**	6.55	**18.61**	20.46
6	25.91	**24.77**	**4.96**	5.72	**4.87**	5.55	**19.51**	22.16
9	**14.13**	17.55	**3.50**	5.22	**3.82**	6.33	**17.49**	21.71
12	**15.02**	17.66	**4.20**	4.69	**4.32**	5.64	**17.52**	19.75
15	**16.67**	17.67	**4.26**	5.04	**4.29**	5.98	**17.84**	20.05
18	**17.18**	19.61	**5.12**	6.38	**4.39**	7.45	**18.36**	21.03
21	**18.50**	21.20	**4.28**	6.47	**4.34**	7.73	**19.28**	20.19
24	**18.31**	21.17	**4.48**	7.86	**4.32**	7.25	**18.86**	21.28
27	**18.81**	21.46	**4.42**	8.39	**4.68**	8.13	**19.87**	20.92
30	**18.83**	22.42	**4.40**	8.82	**5.25**	7.99	**20.13**	21.68

Table 2. Registration accuracy on test samples every 3 epochs for networks trained with physically regularized (Phy) and not physically regularized (Nophy) data generation for the synthetic cubes test cases.

the test sample is measured every 3 epochs.

In the second experiment, we used the "Base + scaling + phy" dataset to evaluate the effect of the number of training samples on the performances of the network. For each training run, only a portion of the dataset was used to train the network, from 0.1% to 100%. The results of this experiment are presented in Fig. 6. Finally, additional results, including qualitative results, are presented in the Supplementary.

3.3. Discussion and conclusion

In Sec. 3.1, the registration accuracy of the network trained on synthetic deformations was evaluated with and without physical regularization.

On the porcine test case from the IHUDeLiver10 dataset, the best registration performance is 2.8 mm, obtained at epoch 24 for the network trained on physically regularized data. However, the accuracy of the network does not improve monotonically during training, suggesting that early stopping may be necessary to obtain the best registration performances on the test set. Additionally, while the physical regularization generally improves performances, it is not true for all epochs. Finally, this experiment would need to be repeated on the full IHUDeLiver10 dataset to better appreciate the registration performances of the method.

On the synthetic cube test case, the difference in accuracy between the networks trained on physically regularized and not physically regularized data is more clear, with the physically regularized method performing almost always better. This may be related to the relative lack of contrast of the synthetic dataset, with each tile of the checkerboard pattern being of constant intensity. Without contrast, the deformation inside the tile can only be inferred from the deformation of the tile edges. With the physical regularization, the network might be able to learn to better interpolate the displacement inside the tile from the displacement at the edges of the tile. Again, while there is no monotonic convergence, the best results are still obtained with the physically regularized data.

In Sec. 3.2, the first ablation study experiment shows the importance of adjusting the synthetic training data distribution to better match the testing data distribution. Despite its simplicity, removing the "scaling" transformation resulted in very poor registration performances, with the network failing to converge. On the other hand, the addition of the biomechanical regularization, which induces a non-negligible additional computational cost, improved the registration performance by a modest amount, and only at some training epochs. However, there are other aspects to take into consideration for physically regularized registration, namely choosing the right biomechanical model with the right parameters, and choosing physically plausible boundary conditions. In our cases, while the random DVF is smooth and diffeomorphic, it does not respect the conservation laws of physics. Due to this, the surface of the organ may be subject to physically implausible deformations, giving rise to unrealistically high strain energy inside of the organ. However, despite these limitations, the best performance is attained by the network trained on the physically regularized dataset, with a clinically relevant accuracy of 2.8 mm (from 6.2 mm before registration).

The second ablation study experiment sheds some light on the number of training samples necessary to learn the 2D-3D registration task. While the accuracy of the networks trained on 18 and 180 samples is never better than 4 mm, the best results are obtained for the network trained with 18000 samples with the networks trained with 1800 to 18000 samples yielding similar performances. Additionally, the results shows that further increasing the size of the dataset would likely lead to diminishing returns, sug-

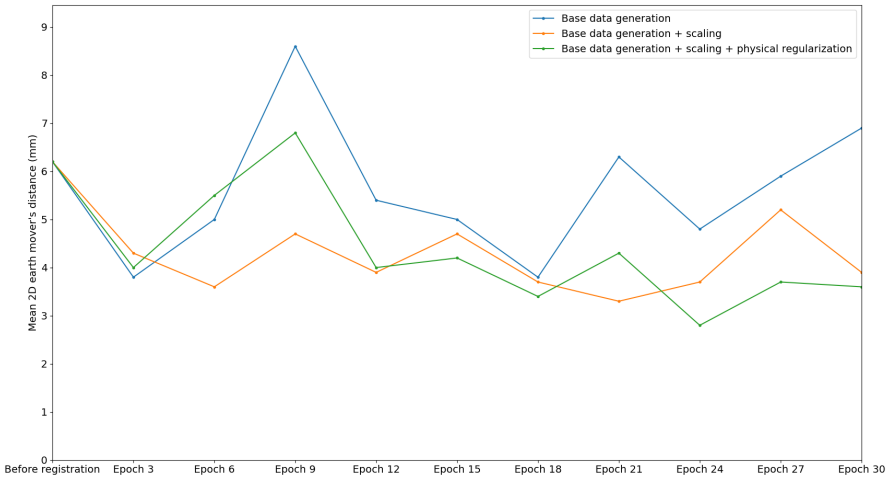

Figure 5. Each of the blue, orange and green curve shows the accuracy of the network every 3 epochs, for different data generation processes. In blue, the accuracy for the dataset generated following Sec. 2.2. In orange, the accuracy for the dataset generated with random scaling of the DVFs. In green, the accuracy for the dataset generated with the scaling and the physical regularization.

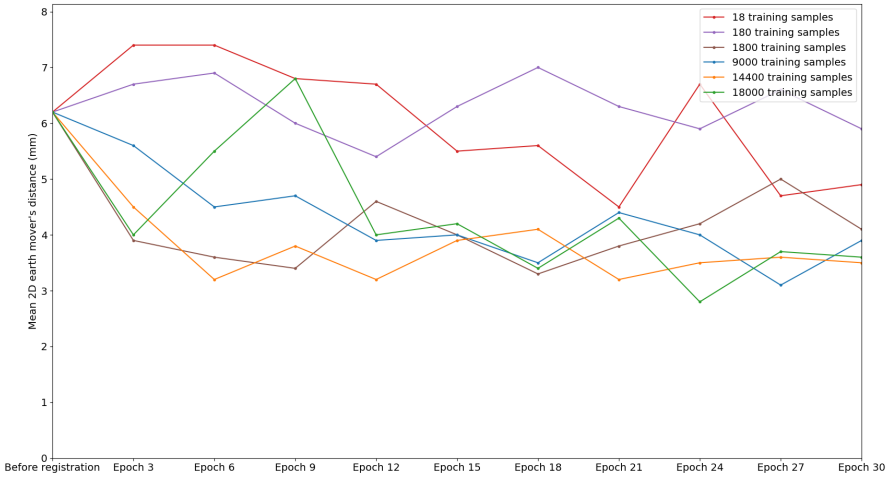

Figure 6. The red, purple, brown, blue, orange and green curves show the accuracy of the network every 3 epochs for dataset sizes of 18, 180, 1800, 14400 and 18000 respectively.

gesting the quality of the network architecture, training and data generation process are more critical aspects for performance. In this work, we have proposed a randomized synthetic data generation methodology for DIR problems, specifically tailored for 2D-3D X-Ray to CT abdominal organ registration, but that is adapted for arbitrary deformations by construction. Additionally, we enforce the physical plausibility of the randomized DVF using a correction step based on a biomechanical model, and have showed in an ablation study that this correction step may allow for better registration performance. We plan on evaluating our methodology in the complete IHUdeLiver10 dataset in future work. **Acknowledgement.** This work was funded by the French national research agency ANR (ANR-20-CE19-0015).

References

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph., 28(3):24, 2009. 5

[2] Stanley Durrleman, Marcel Prastawa, Nicolas Charon, Julie R. Korenberg, Sarang Joshi, Guido Gerig, and Alain Trouvé. Morphometry of anatomical shape complexes with dense deformations and sparse parameters. NeuroImage, 101:35–49, 2014. 2

[3] Markus D. Foote, Blake E. Zimmerman, Amit Sawant, and Sarang C. Joshi. Real-Time 2D-3D Deformable Registration with Deep Learning and Application to Lung Radiotherapy Targeting. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11492 LNCS:265–276, 2019. Publisher: Springer Verlag. 2

[4] Alessa Hering, Stephanie Häger, Jan Moltz, Nikolas Lessmann, Stefan Heldmann, and Bram van Ginneken. CNN-based lung CT registration with multiple anatomical constraints. Med. Image Anal., 72:102139, 2021. 1

[5] Malte Hoffmann, Benjamin Billot, Douglas N. Greve, Juan Eugenio Iglesias, Bruce Fischl, and Adrian V. Dalca. Synthmorph: Learning contrast-invariant registration without acquired images. IEEE Transactions on Medical Imaging, 41(3):543–558, 2022. 1

[6] Julian Krebs, Herve Delingette, Boris Mailhe, Nicholas Ayache, and Tommaso Mansi. Learning a probabilistic model for diffeomorphic registration. IEEE Transactions on Medical Imaging, 38(9):2165–2176, 2019. 1

[7] Francois Lecomte, Valentina Scarponi, Pablo A Alvarez, Juan M Verde, Jean-Louis Dillenseger, Eric Vibert, and Stéphane Cotin. Enhancing fluoroscopy-guided interventions: a neural network to predict vessel deformation without contrast agents. In Hamlyn Symposium on Medical Robotics, pages 75–76, 2023. 4

[8] Megumi Nakao, Mitsuhiro Nakamura, and Tetsuya Matsuda. Image-to-graph convolutional network for 2d/3d deformable model registration of low-contrast organs. IEEE Transactions on Medical Imaging, 41(12):3747–3761, 2022. 2

[9] Thilo Sentker, Frederic Madesta, and René Werner. GDL-FIRE^{4D}: Deep Learning-Based Fast 4D CT Image Registration, pages 765–773. Springer International Publishing, 2018. 1

[10] Hua-Chieh Shao, Jing Wang, Ti Bai, Jaehee Chun, Justin C Park, Steve Jiang, and You Zhang. Real-time liver tumor localization via a single x-ray projection using deep graph neural network-assisted biomechanical modeling. Physics in Medicine & Biology, 67(11):115009, 2022. 1, 2

[11] Hessam Sokooti, Bob de Vos, Floris Berendsen, Boudewijn P. F. Lelieveldt, Ivana Išgum, and Marius Staring. Nonrigid Image Registration Using Multi-scale 3D Convolutional Neural Networks, pages 232–239. Springer International Publishing, 2017. 1

[12] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. Deformable Medical Image Registration: A Survey. IEEE transactions on medical imaging, 2010. 1

[13] Alain Trouve, Mirza Faisal Beg, Michael I Miller, and Laurent Younes. Computing Large Deformation Metric Mappings via Geodesic Flows of Diffeomorphisms. International Journal of Computer Vision, 61(2):139–157, 2005. 2

[14] Mathias Unberath, Jan Nico Zaech, Sing Chun Lee, Bastian Bier, Javad Fotouhi, Mehran Armand, and Nassir Navab. DeepDRR – A Catalyst for Machine Learning in Fluoroscopy-Guided Procedures. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11073 LNCS:98–106, 2018. 4

[15] Mathias Unberath, Cong Gao, Yicheng Hu, Max Judish, Russell H Taylor, Mehran Armand, Robert Grupp, Ka-Wai Kwok, Luigi Manfredi, and Changsheng Li. The Impact of Machine Learning on 2D/3D Registration for Image-Guided Interventions: A Systematic Review and Perspective. Frontiers in Robotics and AI, 8, 2021. ISBN: 2021.716007. 1