Automatic landmark correspondence detection in medical images with an application to deformable image registration

Monika Grewal[a,*] **Jan Wiersma**[b] **Henrike Westerveld**[b] **Peter A. N. Bosman**[a,c,*] **and Tanja Alderliesten**[d]

[a]Centrum Wiskunde and Informatica, Evolutionary Intelligence Research Group, Amsterdam, The Netherlands

[b]University of Amsterdam, Amsterdam University Medical Centers, Location AMC, Department of Radiation Oncology, Amsterdam, The Netherlands

[c]Delft University of Technology, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft, The Netherlands

[d]Leiden University Medical Center, Department of Radiation Oncology, Leiden, The Netherlands

Abstract

Purpose: Deformable image registration (DIR) can benefit from additional guidance using corresponding landmarks in the images. However, the benefits thereof are largely understudied, especially due to the lack of automatic landmark detection methods for three-dimensional (3D) medical images.

Approach: We present a deep convolutional neural network (DCNN), called DCNN-Match, that learns to predict landmark correspondences in 3D images in a self-supervised manner. We trained DCNN-Match on pairs of computed tomography (CT) scans containing simulated deformations. We explored five variants of DCNN-Match that use different loss functions and assessed their effect on the spatial density of predicted landmarks and the associated matching errors. We also tested DCNN-Match variants in combination with the open-source registration software Elastix to assess the impact of predicted landmarks in providing additional guidance to DIR.

Results: We tested our approach on lower abdominal CT scans from cervical cancer patients: 121 pairs containing simulated deformations and 11 pairs demonstrating clinical deformations. The results showed significant improvement in DIR performance when landmark correspondences predicted by DCNN-Match were used in the case of simulated ($p = 0e^0$) as well as clinical deformations ($p = 0.030$). We also observed that the spatial density of the automatic landmarks with respect to the underlying deformation affect the extent of improvement in DIR. Finally, DCNN-Match was found to generalize to magnetic resonance imaging scans without requiring retraining, indicating easy applicability to other datasets.

Conclusions: DCNN-match learns to predict landmark correspondences in 3D medical images in a self-supervised manner, which can improve DIR performance.

 [DOI: 10.1117/1.JMI.10.1.014007]

Keywords: deformable image registration; computed tomography; landmarks detection; deep learning.

Paper 22100GRR received Apr. 19, 2022; accepted for publication Jan. 16, 2023; published online Feb. 25, 2023.

*Address all correspondence to Monika Grewal, monika.grewal@cwi.nl, or Peter Bosman, Peter.Bosman@cwi.nl

1 Introduction

Deformable image registration (DIR) is a task of aligning a source (or moving) image to a target (or fixed) image by optimizing a displacement vector field. The aligned source image can then be computed by resampling the source image at the spatial locations specified by the mapping. DIR has tremendous application possibilities in the radiation treatment workflow required for cancer treatment e.g., automatic contour propagation,[1,2] dose accumulation.[3–5] However, DIR in regions such as the pelvis is challenging due to large local deformations and appearance differences caused by physical processes such as bladder filling, and the presence of gas pockets and contrast agents.[2] In such DIR scenarios, the existing non-linear intensity-based registration approaches[6–8] often get stuck in a local minimum.[4] Many previous studies[9–14] have shown that landmark correspondences between the images to be registered can provide additional guidance to the intensity-based DIR methods and help overcome local minima. However, to the best of our knowledge, such an approach has not been tested on pelvic scans.

Manual annotation of landmarks for DIR in the clinic is not practically tractable due to two main reasons. First, a high number of landmarks is desired, and it is difficult to unambiguously define such a high number of landmarks manually. Second, manual annotations require lots of time from clinicians, which is hardly available. Therefore, an automatic method for finding landmark correspondences is required. Although many endeavors have been made in the direction of automatic landmarks correspondence detection in medical images,[14–16] there remain significant gaps to fill. The existing methods usually employ large pipelines consisting of multiple components, each component using multiple hyperparameters derived from image features specific to the underlying dataset. Consequently, the entire pipeline is sensitive to small variations in local image intensities and choices of hyperparameters, making application to a new dataset difficult. Moreover, in datasets such as pelvic scans with ill-defined boundaries between soft tissues, intensity gradient–based landmark detection may not work at all.

Convolutional neural networks (CNNs) are known to learn deep features from images, which are robust to small variations in local image intensities. In recent years, deep CNNs have not only shown remarkable performance in difficult computer vision tasks in medical imaging,[17,18] but also good generalization to unseen data. Moreover, with the advances in the available computational resources, CNN-based solutions turn out to be faster than their traditional counterparts. Therefore, there is a strong motivation to replace the entire pipeline for automatically finding landmark correspondences by a deep CNN. Recently, some deep CNN methods have been developed for automatic landmark detection in medical images,[19–21] but these are limited to either 2D datasets or supervised learning of a few manually annotated landmarks. Other relevant works include methods for landmark propagation from a template image by learning pixel-wise anatomical embeddings[22] or through DIR.[23] While such methods allow for single shot landmark detection in a new image, the requirement of manual annotation of landmarks on the template image still exists. Another study uses unsupervised image registration as a proxy task to discover landmarks shape descriptors,[24] but this method is limited to discovering a small number of landmarks (∼100 landmarks per image pair).

In this study, we present a deep CNN (referred to as DCNN-Match) for automatic landmarks correspondence detection (i.e., simultaneous landmark detection as well as matching) in 3D images. The presented method is an extension of our method for 2D images.[21] Briefly, the neural network is trained on pairs of lower abdominal 3D computed tomography (CT) scans such that the network learns to predict landmarks at salient locations in both the images along with the correspondence score of each landmark pair. One key feature of the presented method is that unlike supervised methods, the neural network in the presented method is trained in a self-supervised manner without using any manual annotations. This is important because manual annotations on medical images are not always readily available, mainly because it is time-consuming to create them.

It is essential to investigate the added value of automatic landmarks correspondence detection toward the improvement of the DIR solutions to estimate the potential deployability of landmarks-guided DIR approaches in the clinic. Existing studies have investigated the added value of automatic landmark correspondences towards DIR independently of the underlying automatic landmark detection method.[10,11,14] Since change in the automatic landmarks

correspondence detection method changes the aspects of the automatic landmarks, e.g., spatial distribution and matching accuracy, the effect of the automatic landmarks on the DIR performance is likely to be affected as well. Therefore, we believe that developing a method for automatic landmarks correspondence detection and at the same time integrating it with a DIR pipeline can provide numerous insights. To this end, we have integrated our method for automatic landmark detection and matching with an existing DIR software so that the added value of using landmark correspondences in solving DIR problems can be assessed. Further, we investigate five different variants of the developed method by use of different loss functions during training that each predict landmark correspondences with different spatial distributions and matching errors, to assess the effect of different types of automatic landmark correspondences towards the improvement of DIR. The present work has the following contributions:

1. We extended our previously published end-to-end self-supervised deep-learning method for automatically finding landmark correspondences in medical images from 2D to 3D. The key highlights of the method are:
   a. the method does not set any prior on the definition of landmarks and
   b. the method does not require manual annotations for training
2. We integrated our automatic landmark correspondence detection method in 3D (DCNN-Match) with an open-source registration software Elastix[6,25] to develop a DIR pipeline that utilizes additional guidance information from automatic landmark correspondences. We used this DIR pipeline to investigate the added value of automatic landmark correspondences in providing additional guidance to the DIR method and finding better DIR solutions.
3. We varied the landmarks correspondence detection method and investigated how it affected the added value to the DIR method. We explored five different variants of the proposed automatic landmarks correspondence method.
4. We experimentally investigated the generalization capability of our proposed automatic landmarks correspondence detection method to a magnetic resonance imaging (MRI) dataset.

2 Materials and Methods

In the following sections, we describe DCNN-Match (Sec. 2.1), and the DIR pipeline that uses the information from automatic landmark correspondences predicted by DCNN-Match to guide the registration (Sec. 2.2). Sections 2.3 and 2.4 provide details of implementation and hyperparameters for reproducibility. Sections 2.5, 2.6, 2.7, and 2.8 describe the datasets, experiments, evaluation metrics, and statistical testing used in the experiments, respectively.

2.1 *DCNN-Match*

We extended our approach[21] for finding landmark correspondences in 2D CT scan slices to work on 3D CT scans. The different components of the 3D approach are shown in Fig. 1. Briefly, the approach proposed in Ref. 21 consists of a Siamese network with three modules: (a) two CNN branches with shared weights, (b) a sampling layer, and (c) a descriptor matching module. The CNN branches comprise an image-to-image translation network that maps an input image to a feature map. The architecture of the network is derived from the famous UNet architecture[26] proposed for image segmentation. For a given pair of target image (I_{target}) and source image (I_{source}), the CNN branches predict a landmark probability map describing the probability $\hat{p}_i^{I_x}$ ($x \in \{\text{target}, \text{source}\}$) of each spatial location i being a landmark. The sampling layer is a parameter-free module that samples K (hyperparameter) landmark locations with top landmark probabilities during training. During inference, the sampling layer samples all landmark locations with landmark probabilities above a threshold. We used the value 0.5, same as in Ref. 21.

Additionally, the sampling layer samples a feature vector from the feature maps of the last two downsampling levels in the CNN branch at the coordinates of each i'th landmark location

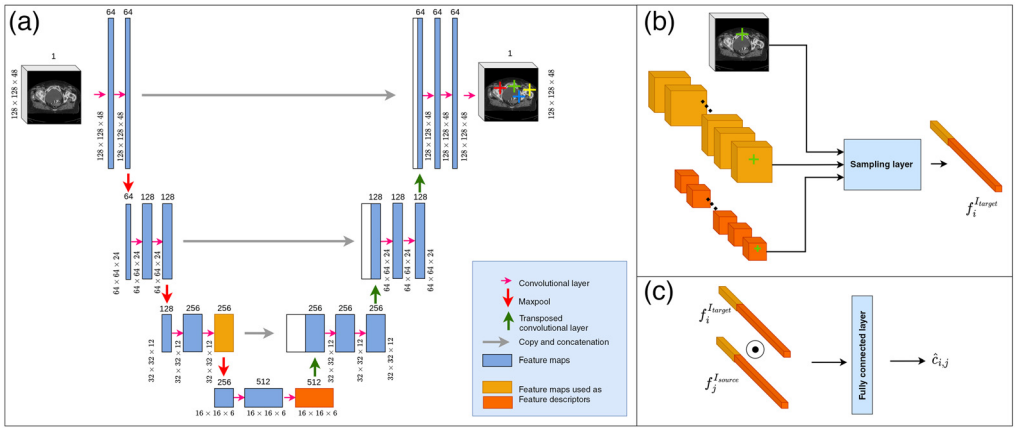

Fig. 1 Illustration of the components of DCNN-Match. (a) Illustration of different layers in the shared CNN branch used for landmark detection and feature description. (b) The sampling layer samples the feature maps of the last two downsampling levels in the CNN branch at the locations described by the landmark probability map. (c) The descriptor matching module realized by a fully connected layer predicts the matching probability of a feature descriptor pair.

and constructs the feature descriptor $f_i^{I_x}$ by concatenating the sampled feature vectors. This allows for efficient use of the network weights by simultaneous learning the landmark detection as well as feature description of each landmark without unnecessarily increasing the network size. Moreover, the concatenation of features from different downsampling levels emulates the behavior of multi-scale feature description, which otherwise, is achieved by calculating features from a Gaussian pyramid representation of the image. Following the calculation of feature descriptors for each landmark location, the sampling layer creates feature descriptor pairs $(f_i^{I_{\text{target}}}, f_j^{I_{\text{source}}})\ \forall\ i = 1, 2, \ldots, K$ in I_{target} and $\forall\ j = 1, 2, \ldots, K$ in I_{source} to feed to the descriptor matching module. The descriptor matching module predicts the landmark matching probabilities corresponding to each feature descriptor pair.

2.1.1 *Self-supervised training*

The network is trained in a self-supervised manner on pairs of target (I_{target}) and source (I_{source}) lower abdominal CT scans containing simulated deformations. The details on the generation of target and source image pairs are provided in Sec. 2.3.

Following the sampling of landmark locations $i = 1, 2, \ldots, K$ in I_{target} and $j = 1, 2, \ldots, K$ in I_{source} along with their corresponding feature descriptors $f_i^{I_{\text{target}}}$ and $f_j^{I_{\text{source}}}$, feature descriptor pairs $(f_i^{I_{\text{target}}}, f_j^{I_{\text{source}}})$ are constructed in the sampling layer. The feature descriptor pairs are considered corresponding to all i and j, allowing for feature descriptor matching between far-away locations in the images without requiring encoding of the underlying deformation field explicitly. Since the simulated deformations used to create source and target image pairs during training can not represent the complex large deformations in a clinical setup exactly, learning the feature descriptor matching not explicitly dependent on the underlying deformation field is likely to help the neural network generalize better to clinical scenario.

The ground truth $c_{i,j}$ of the correspondence of each feature descriptor pair is calculated on-the-fly based on the known simulated deformation. Each sampled landmark location in the target image is projected onto the source image based on the known simulated deformation and the nearest predicted landmark (within a distance of 2 voxels = 4 mm) in the source image is considered its match. We used a threshold of 4 mm (instead of image resolution = 2 mm) to find a reasonable number of landmark matches from random predictions in the beginning of the training to ensure sufficient supervision. The value of $c_{i,j} = 1$ for matching and $c_{i,j} = 0$ for non-matching feature descriptor pairs. Subsequently, the ground truth $p_i^{I_x}$ for the landmark probability of landmark location i in image $I_x, x \in \{\text{target}, \text{source}\}$ is determined as follows:

$$p_i^{I_x} = \begin{cases} 1 & \text{if } \exists!\, j \in \{0, 1, 2, \ldots, K\} \text{ in image } I_y, y \in \{\text{target}, \text{source}\},\ y! = x \wedge c_{i,j} = 1 \\ 0 & \text{otherwise} \end{cases}. \tag{1}$$

The ground truths $c_{i,j}$ are used directly as ground truths for the matching probability of the feature descriptor pairs $(f_i^{I_{\text{target}}}, f_j^{I_{\text{source}}})$. In other words, the ground truth is generated such that the landmark probability as well as the descriptor matching probability is high for the matching locations between the two images and low otherwise. The network is trained by minimizing a multi-task loss defined as follows:

$$\begin{aligned} \text{Loss} = {} & \text{LandmarkProbabilityLoss}_{I_{\text{target}}} + \text{LandmarkProbabilityLoss}_{I_{\text{source}}} \\ & + \text{DescriptorMatchingLoss}. \end{aligned} \tag{2}$$

The $\text{LandmarkProbabilityLoss}_{I_x}$ for the probabilities of landmarks in image $I_x, x \in \{\text{target}, \text{source}\}$ is defined as

$$\text{LandmarkProbabilityLoss}_{I_x} = \frac{1}{K} \sum_{i=1}^{K} ((1 - \hat{p}_i^{I_x}) + \text{CrossEntropyLoss}(\hat{p}_i^{I_x}, p_i^{I_x})), \tag{3}$$

where CrossEntropyLoss is the cross entropy loss between predicted landmark probability $\hat{p}_i^{I_x}$ and ground truth $p_i^{I_x}$ of the i'th sampled location. K is the total number of sampled landmark locations in image I_x. Further details of the LandmarkProbabilityLoss are omitted for brevity and can be found in Ref. 21.

The DescriptorMatchingLoss allows the network to learn feature descriptor matching automatically and is defined as follows:

$$\text{DescriptorMatchingLoss} = \text{DescriptorHingeLoss} + \text{DescriptorCELoss}. \tag{4}$$

DescriptorHingeLoss is defined as follows:

$$\begin{aligned} \text{DescriptorHingeLoss} = \sum_{i=1,j=1}^{K,K} \Bigg(& \frac{c_{i,j} \max(0, \| f_i^{I_{\text{target}}} - f_j^{I_{\text{source}}} \|^2 - m_{\text{pos}})}{K_{\text{pos}}} \\ & + \frac{(1 - c_{i,j}) \max(0, m_{\text{neg}} - \| f_i^{I_{\text{target}}} - f_j^{I_{\text{source}}} \|^2)}{K_{\text{neg}}} \Bigg), \end{aligned} \tag{5}$$

where $f_i^{I_{\text{target}}}$ and $f_j^{I_{\text{source}}}$ are the feature descriptors corresponding to the i'th and j'th landmark locations in the input images I_{target} and I_{source}, respectively; $c_{i,j}$ is the ground truth matching probability for the feature descriptor pair $(f_i^{I_{\text{target}}}, f_j^{I_{\text{source}}})$; m_{pos} and m_{neg} are the margins for the $L2$-norm of matching (positive class) and non-matching (negative class) feature descriptor pairs. The Hinge losses corresponding to positive and negative classes are normalized by K_{pos} (number of positive feature descriptor pairs) and K_{neg} (number of negative feature descriptor pairs), respectively to account for the class imbalance between positive and negative feature descriptor matches. DescriptorCELoss is defined as follows:

$$\text{DescriptorCELoss} = \sum_{i=1,j=1}^{K,K} \left(\frac{\text{WeightedCrossEntropy}(\hat{c}_{i,j}, c_{i,j})}{(K_{\text{pos}} + K_{\text{neg}})} \right), \tag{6}$$

where $\hat{c}_{i,j}$ is the predicted matching probability; WeightedCrossEntropy represents the binary cross entropy loss where the loss corresponding to the positive class is weighted by the frequency of negative examples and vice versa.

In the beginning of the training, the predicted landmark probability maps by the CNN branches are random and by chance only a few landmark locations have correct correspondence

(i.e., $c_{i,j} = 1$) between images. The loss defined in (2) encourages high landmark probability at these locations as well as high feature descriptor matching probability for the feature descriptor pairs of these locations and low landmark probability and feature descriptor matching probability otherwise. Additionally, the term $(1 - \hat{p}_i^{I^s})$ in (3) encourages high landmark probability at all locations, i.e., encourages more landmark locations to have correct correspondence in the other image. Consequently as the training progresses, the network learns to identify salient locations in the images that have correct correspondence in the other image as well and predicts high landmark probabilities at the these locations.

2.1.2 *End-to-end*

The conventional approach to establish landmark correspondences between an image pair utilizes the following steps:

1. Landmark detection, in which landmarks are detected in both the images independently.
2. Feature description, wherein a vector (often called "descriptor") is calculated to describe the image properties surrounding the landmark location. An example of a feature descriptor is scale invariant feature transform,[27] which calculates the histograms of orientations from the image patches of different scales around the landmark.
3. Landmark matching, wherein landmark descriptors in both the images are matched using a matching algorithm. A straightforward matching algorithm is brute force matching, which aims at finding the best match among all the landmark locations in the source image for each landmark location in the target image.

Our approach replaces each of the abovementioned components with a neural network module, and connects the neural network modules such that the gradients flow from the end to the inputs. The modules of landmark detection and description are represented by the CNN branches of the Siamese network. The task of landmark matching is performed by the descriptor matching module. It is important to mention that the key feature of DCNN-Match lies in the assembling of different modules to provide a simple, end-to-end deep-learning solution for simultaneous landmark detection, description, and matching automatically. Therefore, the proposed approach can be easily modified, e.g., it may be improved by the use of a different neural network in any of the modules.

2.1.3 *Extension to 3D images*

We have extended our original approach proposed in Ref. 21 to work on 3D images by performing three modifications. The first obvious modification was to use 3D convolutional kernels (kernel size = $3 \times 3 \times 3$) instead of 2D convolutional kernels in the CNN branches. The sampling layer and the feature descriptor matching module were also adapted for 5D tensors arising from training on 3D images. The generation of a valid mask during training as described in[21] Sec. 2.4 was also adapted for 3D images. The valid mask makes the network learn a content-based prior to predict landmarks only in the regions that include patient anatomy and not in the background or the CT couch.

Second, since we had a considerably larger training dataset (details in Sec. 2.5) as opposed to Ref. 21, we kept the same number of kernels in each layer as the original UNet architecture.[26] Third, we trained the network on 3D patches of the entire CT due to GPU memory constraints. During inference, we evaluated the network on the patches belonging to the same spatial locations in the target and source images. The patches were cut with 50% overlap and the final output combined the predicted landmark pairs in all patches. All the corresponding landmarks predicted in all the overlapping patches were considered landmarks. Using a small patch size restricts the network from learning landmark matches in locations that are far apart in the two images. Therefore, the patch size has to be decided while keeping in mind the spatial extent of deformations we want the network to learn. This is further described in the hyperparameters Sec. 2.4.

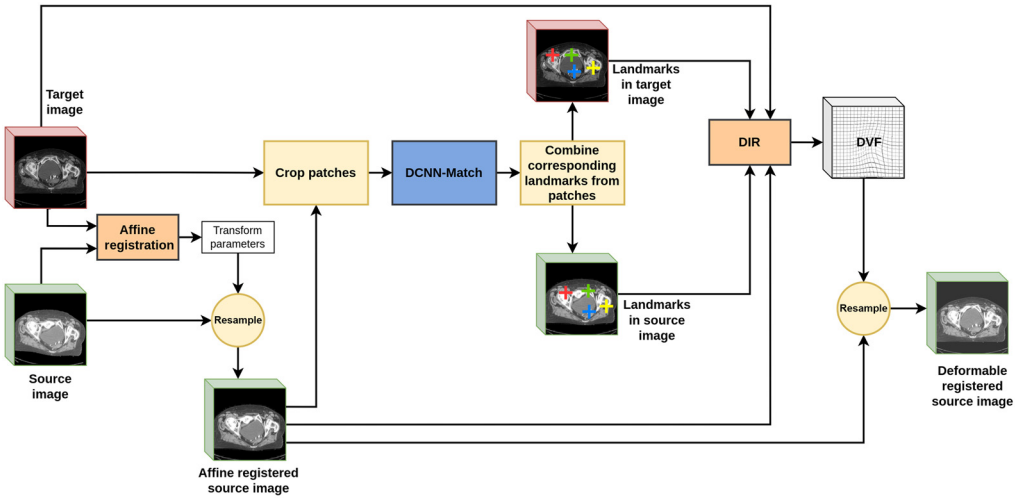

Fig. 2 DIR pipeline with automatic landmarks correspondence detection using DCNN-Match. The source image is affine registered with the target image followed by automatic landmarks correspondence detection using DCNN-Match. DCNN-Match provides the locations of corresponding landmarks (shown with similar colored cross-hairs) in both the target and affine registered source image. The DIR module finds a DVF by utilizing the additional guidance information from automatic landmark correspondences. The final transformed (deformable registered) source image is obtained by resampling the affine registered source image according to the obtained DVF.

2.2 *DIR with Additional Guidance from Automatic Landmark Correspondences*

We integrated DCNN-Match with the open-source registration software Elastix[6,25,28] to create a pipeline for DIR that utilizes the additional guidance information from automatic landmark correspondences. A schematic of the DIR pipeline is provided in Fig. 2.

DIR requires calculation of a DVF that maps each spatial location in the target image to a spatial location in the source image. In Elastix, the DVF is parameterized by B-splines and the coefficients of B-splines are optimized by non-linear optimization. We align the source CT scans with the target CT scans using affine registration before performing DIR. The parameters of the 3D affine transformation matrix (i.e., translation, rotation, scale, and shear) are optimized by maximizing the normalized mutual information between the target and source scans. The target and the affine registered source CT scan are input to the DCNN-Match, which provides the locations of corresponding landmarks in both the scans. The DIR module in Elastix takes the target image, affine registered source image, and the pairs of corresponding landmarks in both the images as input. The DIR is performed by optimizing the following objective function:

$$\begin{aligned} f_{\text{Guidance}} = {} & \text{weight}_0 \text{AdvancedMattesMutualInformation} \\ & + \text{weight}_1 \text{TransformBendingEnergyPenalty} \\ & + \text{weight}_2 \text{CorrespondingPointsEuclideanDistanceMetric}. \end{aligned} \tag{7}$$

where AdvancedMattesMutualInformation represents the maximization of mutual information between two scans (for details refer to Ref. 29), TransformBendingEnergyPenalty is a regularization term that penalizes large transformations, and CorrespondingPointsEuclideanDistanceMetric is used for minimizing the Euclidean distance between the landmarks in the target CT and the landmarks in the source CT. weight_0, weight_1, and weight_2 control the relative contribution of each term towards the objective function.

2.3 *Implementation*

The DIR pipeline was developed in Python. We used the PyTorch framework[30] for developing DCNN-Match. The training was done on an RTX 2080 Ti GPU and took approximately 21 h.

The weights of DCNN-Match were initialized using the He norm method.[31] The training was done using the Adam optimizer[32] with a learning rate of $1e^{-4}$. The neural network weights were regularized using a weight decay of $1e^{-4}$.

We randomly cropped 3D patches of dimension $128 \times 128 \times 48$ from the entire CT scan volume and used them as target images. The source images were generated on-the-fly by applying one of the following random transformations on the target images: translation, rotation, scale, or elastic transformations. The magnitudes of the affine transformations along all axes were sampled from the following uniform distributions: $U(-12\text{ mm}, 12\text{ mm})$, $U(-20\text{ deg}, 20\text{ deg})$, and $U(0.9, 1.1)$ for translation, rotation, and scale, respectively. The elastic transformations were applied so as to simulate the two types of soft tissue deformations present in the lower abdominal scans: (a) large local deformations, e.g., bladder filling, and (b) small tissue deformations everywhere in the image. The large local deformations were simulated by a 3D Gaussian DVF ($\text{DVF}_{\text{large}}$) of magnitude at center $= U(2\text{ mm}, 24\text{ mm})$ and $\sigma = U(64\text{ mm}, 128\text{ mm})$ at a random location in the image. The small deformations everywhere in the image were simulated by Gaussian smoothing of a random DVF ($\text{DVF}_{\text{small}} = U(1\text{ mm}, 12\text{ mm})$) at each location. $\text{DVF}_{\text{large}}$ and $\text{DVF}_{\text{small}}$ were additively applied to the target image to generate the source image with elastic transformation.

2.4 *Hyperparameters*

Apart from the conventional hyperparameters involved in designing and training a DCNN e.g., network depth and width, optimizer, and learning rate, there are two hyperparameters specific to DCNN-Match: patch dimensions and the number of sampling points during training (K). As indicated in the previous section, we used a patch size of $128 \times 128 \times 48$ ($256\text{ mm} \times 256\text{ mm} \times 96\text{ mm}$). This way the neural network's field-of-view (FOV) was maximum given the network depth and GPU memory constraints, which ensured that the landmark correspondences could be learned for deformations as large as 128 mm in-plane and 48 mm along the transverse axis. Similar to Ref. 21, $K = 512$ was used based on the visual inspection that the predicted landmarks in the validation set (details in Sec. 2.5.1) covered the image sufficiently.

In Elastix, we used the advanced mattes mutual information as a similarity metric because it has been found successful in earlier studies on DIR.[2] For deciding other hyperparameters, such as the number of iterations, step size, step decay, weight_0, weight_1, and weight_2, we used the development set (details in Sec. 2.5.1). For this purpose, the pairs of target and source images were generated in a manner similar to the training set. About 100 locations were sampled randomly on the target image and their corresponding location in the source image was established by transforming the coordinates with the inverse DVF used for generating the source image. The hyperparameters were tuned based on the following observations on the development set: the transformed source image after registration was not distorted and showed no visible folding, the image alignment at the 100 randomly sampled locations improved after registration. The exact configuration of Elastix used for affine registration and DIR is provided in Appendix A.

2.5 *Data*

An overview of the data is provided in Fig. 3. We retrospectively included the CT and MRI scans from female patients (age range 22 to 95 years), who received radiation treatment in the lower abdominal region between the year 2009 and 2019 at the Amsterdam University Medical Centers, location AMC, the Netherlands. The data were transferred in anonymized form through a data transfer agreement. A subset of these scans was the same as used in a previous study.[21]

2.5.1 *Training and validation set*

A total of 1671 CT scans of 831 patients were used for developing the approach: 1335 CT scans for training and 336 CT scans for validation. A subset containing 10 CT scans from the validation set (referred to as the development set) was used to tune the hyperparameters of the

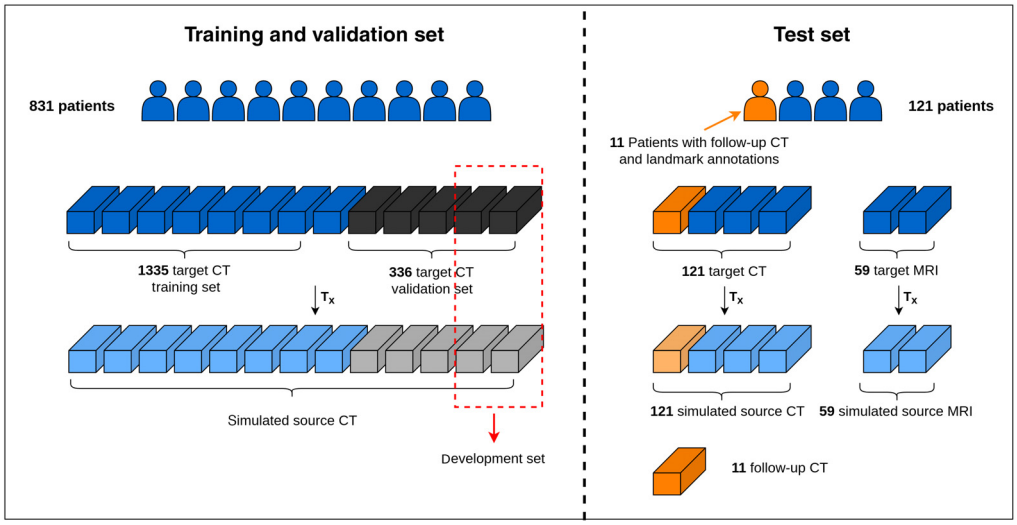

Fig. 3 Data overview. The vertical dashed gray line depicts the patient-level split between the training and validation set, and test set.

DIR pipeline. All the CT scans were resampled to have 2 mm × 2 mm × 2 mm voxel spacing and the image intensities were converted from the Hounsfield units to a range of 0 to 1 after windowing.

2.5.2 *Simulated deformations test set – CT*

We tested the performance of DCNN-Match and the DIR pipeline on a curated dataset of 121 CT scans belonging to 121 patients, who received radiation treatment for cervical cancer. The mean FOV of acquisition of the CT scans was 546 mm × 546 mm × 368 mm and the scans were resampled to 2 mm × 2 mm × 2 mm voxel spacing. The available CT scans were used as target images and corresponding source images were simulated by applying random elastic transformations to the target CT scans according to the method described in Sec. 2.3 above. Further, an example of the simulated deformation and the obtained source CT is shown in Fig. 4(a).

In each pair of target and source image, 100 corresponding locations were sampled with uniform random distribution. These sampled locations were used as validation landmarks for assessing the performance of DCNN-Match and the DIR pipeline.

2.5.3 *Clinical deformations test set – CT*

The CT scans in a clinical setup exhibit complex biomechanical deformations, including discontinuities in the deformation field around sliding tissues and large deformations that may not be Gaussian. The random Gaussian DVF used for deforming the images to obtain a simulated test set is an oversimplification of the underlying situation. Therefore, it is essential to investigate if the observations on the simulated deformations test set hold in the clinical setting as well. To this end, additional CT scans (referred to as follow-up scans) were searched in the clinical database for a subset of patients in the test set (11 patients). The first CT scans from these patients were used as target images and the corresponding follow-up CT scans were used as source images.

Corresponding landmarks at 29 locations were manually identified in each target and source CT scan by a clinical expert. These landmarks included six fiducial markers in the vaginal wall, and anatomical landmarks, e.g., aortic bifurcation, cervical os, and os coccygis. Since clinically available scans were used, the number of fiducial markers were different in each patient in accordance with the treatments given to the patients. The majority of the patients' scans had three fiducial markers, while some had less or more. If a patient's scan had less than three fiducial markers, calcification (if present) in corresponding anatomical locations were used as landmarks. If a patient's scan had more than three fiducial markers, only three of them were used. An

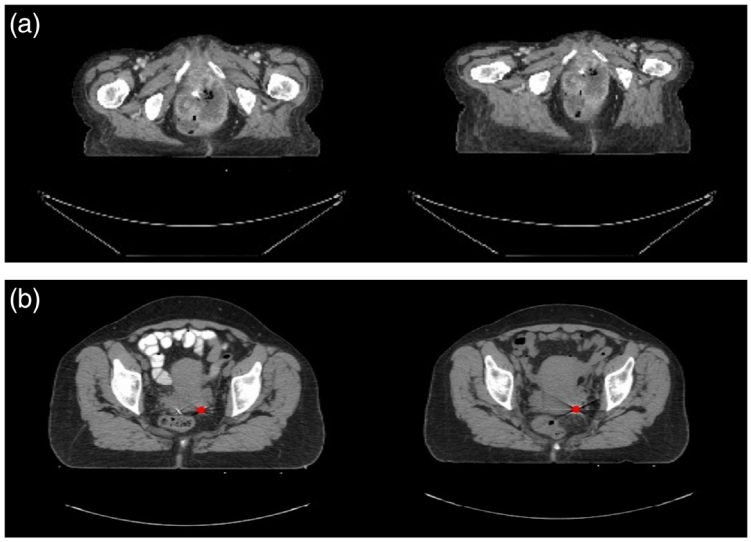

Fig. 4 Transverse slices from representative examples. (a) Simulated deformations test set: the source CT (right) is obtained by applying an elastic transformation to the target CT (left). (b) Clinical deformations test set: the landmark at the location of a fiducial marker (shown with red dot) in the target (left) and source (right) CT is shown. Note the appearance difference in the bowel due to contrast.

example landmark location is shown in Fig. 4(b) and the complete list of landmark locations is provided in Appendix B.

2.5.4 *Simulated deformations test set: MRI*

MRI scans of 59 cervical cancer patients (subset of the 121 cervical cancer patients mentioned in Sec. 2.5.2, who received brachytherapy treatment) acquired during brachytherapy treatment delivery were used to investigate the generalization capability of DCNN-Match. The mean FOV of acquisition of the MRI scans was 199 mm × 199 mm × 152 mm and the scans were resampled to 2 mm × 2 mm × 2 mm voxel spacing. The pairs of source and target scans were generated in a similar way to the CT scans (Sec. 2.5.2).

2.6 *Experiments*

We conducted three types of experiments. The first type of experiments were aimed to gain insights in the working of DCNN-Match by changing the DescriptorMatchingLoss (Secs. 2.6.1 and 2.6.2). The second type of experiments were done to investigate the effect of automatically predicted landmark correspondences on the performance of DIR (Sec. 2.6.3). We also investigated how the changes in DescriptorMatchingLoss affected the added value of the automatic landmark correspondences toward the performance of DIR. Third, we investigated the generalization capability of DCNN-Match on a different modality (Sec. 2.6.4).

2.6.1 *Descriptor loss*

We trained three versions of DCNN-Match, each with a different DescriptorMatchingLoss. The first version was trained with only the DescriptorHingeLoss defined in Eq. (5). This version is referred to as DCNN-Match Hinge. DCNN-Match Hinge was trained with $m_{\text{pos}} = 0$ and $m_{\text{neg}} = 1$. In the second version, only DescriptorCELoss Eq. (6) was employed. We refer to this version as DCNN-Match CE. Next, we trained the network with a linear combination of DescriptorHingeLoss and DescriptorCELoss Eq. (4), which is referred to as DCNN-Match Hinge + CE.

2.6.2 *Positive margin in the hinge loss*

We considered that the $L2$-norm of the descriptor pairs of highly deformed regions would be high, and these pairs would be difficult to match. Further, it is intuitive to think that the landmark matches in regions of high deformation would provide more added value to the DIR approach. To allow the network to focus more on matching these pairs, we trained DCNN-Match Hinge + CE with two values for m_{pos}: 0.1 and 0.2. These versions are referred to as DCNN-Match Hinge0.1 + CE and DCNN-Match Hinge0.2 + CE, respectively. The value of $m_{\text{pos}} > 0$ in the DescriptorHingeLoss makes the loss term 0 for descriptor pairs whose $L2$-norm is $<m_{\text{pos}}$, i.e., the network already identifies the descriptor pairs as matching. Thus, the gradients are influenced only by the descriptor pairs which are difficult to match. Consequently, the network should be able to predict difficult landmark correspondences in the highly deformed regions accurately.

2.6.3 *Effect of additional guidance from automatic landmark correspondences*

To assess the effect of additional guidance from automatic landmark correspondences on the DIR, we compared the results from the DIR pipeline with [$\text{weight}_2 = 0.01$ in Eq. (7) as obtained from hyperparameter tuning on the development set] and without [$\text{weight}_2 = 0$ in Eq. (7)] automatic landmarks correspondence detection.

2.6.4 *Generalization to MRI dataset*

Given the capability of deep neural networks to learn robust features, and the self-supervised nature of our training approach, optimistically one would expect that the developed approach would generalize to different datasets. To this end, we tested DCNN-Match on pairs of MRI scans containing simulated deformations (described in Sec. 2.5.4) without retraining. Compared to the training set, the MRI scans were not only different in imaging modality, but also in the FOV of acquisition.

2.7 *Evaluation*

2.7.1 *Spatial matching errors of landmark correspondences*

In the simulated deformations test set, the landmarks on the source CT scans were projected on the target CT scans using the known transformation between them. The Euclidean distances between the landmarks on the target CT scans and the projection of their corresponding landmarks predicted by the network were calculated. The Euclidean distance gives a measure of the spatial matching error of the predicted landmark correspondences. The spatial matching errors were compared between all versions of DCNN-Match.

Quantitative analysis of the spatial matching errors of the predicted landmark correspondences is not feasible in the clinical deformations test set due to the absence of the ground truth DVF. To provide some insights into the performance on the clinical deformations test set, we conducted a validation study on a subset of the data. For this purpose, we randomly sampled 75 predicted landmarks from DCNN-Match CE in two patients (total 150 landmark correspondences). A radiation oncologist (henceforth, referred to as clinician) ranked these landmark correspondences on a 3 point Likert scale: 1 = good match (roughly within 5 mm distance), 2 = moderate match (roughly within 10 mm distance), 3 = poor or wrong match (roughly more than 15 mm distance) in a 3D (axial, sagittal, and coronal) image viewer. The (approximate) spatial matching errors were calculated based on the ranking provided by the clinician. The clinician also labelled the anatomical location of the landmarks in target CT scans according to the following categories: (a) bony anatomy, (b) soft tissue (i.e., muscles, fatty tissue, and fascia), (c) bowel, i.e., large and small bowel, including gas pockets, and (d) other (including organs and blood vessels i.e., veins and arteries). We also analyzed the spatial matching errors of the predicted landmark correspondences separately for each anatomical category.

2.7.2 *Target registration error*

In the clinical deformations test set, we transformed the manually annotated landmarks in the target images according to the estimated DVF after DIR using the transformix module in SimpleElastix[28] (documentation on using transformix in SimpleElastix can be found at SimpleElastix documentation and Elastix manual). We calculated their Euclidean distance with the corresponding landmarks in the source image. This measure is often referred to as target registration error (TRE). We calculated the TRE values after initial affine registration and before the DIR (TRE_{before}) and after DIR (TRE_{after}) for all experiments. In the simulated deformations test set, TRE calculations were done using the randomly sampled validation landmarks described in Sec. 2.5.

It should be noted that the TRE calculations were done in image space, i.e., the landmarks were represented by the center of a voxel. We chose this setup because the automatic landmarks are predicted in image space. However, this setup may give rise to discretized TRE values.

2.7.3 *Landmark correspondences vs. underlying deformation*

It is intuitive to think that the DIR performance in a specific region is dependent on the underlying deformation in that region. Concordantly, the distribution of landmarks with respect to the underlying deformation would impact the additional guidance provided by the landmarks overall. Therefore, it is important to investigate the choice of landmark locations by the network with respect to the extent of deformation at those locations. To this end, we partitioned the voxels in the source images in the simulated deformations test set into bins of different deformations. For each DCNN-Match variant, the spatial density of predicted landmark correspondences was calculated in each bin of the underlying deformation by dividing the number of landmarks with the number of voxels in each bin.

Similarly, we calculated the percentage of automatic landmarks below 4 mm spatial matching errors (as a surrogate for landmarks correspondence accuracy) and TRE values of validation landmarks (as a measure of DIR performance) in each deformation region. The threshold of 4 mm was chosen because the same threshold was used during training. In each deformation region, we analyzed the TRE values in light of the spatial density and landmarks correspondence accuracy to gain insights about what aspects of automatic landmarks affect the DIR performance.

2.7.4 *Determinant of spatial Jacobian*

Evaluating the performance of DIR is a difficult task and TRE can only give an estimate of performance on sparse image locations. Moreover, TRE can give a biased perspective of the DIR performance because of the observer subjectivity in the manual annotation of landmark locations. In order to assess whether the obtained DVF is anatomically plausible or not, the determinant of the spatial Jacobians of the DVF is a good measure. The negative values in the determinant of the spatial Jacobian represent singularities in the DVF and indicate image folding in those regions. Therefore, we also investigated the determinant of the spatial Jacobians of the obtained DVFs after DIR.

2.8 *Statistical Testing*

The statistical testing was done using IBM SPSS Statistics for Ubuntu (Version 27.0, IBM Corp. Released 2020. Armonk, NY: IBM Corp).[33] We tested the null hypothesis that the TRE_{after} values in the test sets were the same in the following experimental scenarios: DIR without additional guidance from corresponding landmarks, and DIR with additional guidance from five different variants of DCNN-Match.

Kolmogorov–Smirnov tests for normality revealed that the TRE_{after} values were not normally distributed in any of the experimental scenarios. Therefore, we used the related samples Friedman's two way analysis of variance by Ranks test followed by post-hoc pairwise comparisons using Dunn–Bonferroni test.[34] An alpha of 0.05 with Bonferroni correction for multiple comparisons was considered significant.

3 Results

The average inference time of DCNN-Match variants for predicting landmark correspondences in one CT scan pair was ∼20 s. A representative example of predicted landmark correspondences is shown in Fig. 5. The images in the figure are shown with the couch table cropped for better visualization, but the automatic landmark correspondence detection as well as DIR were performed on full CT scans without any cropping.

3.1 *Number of Landmark Correspondences*

The number of landmark correspondences predicted per image on the simulated test set and clinical test set is given in Table 1 and Fig. 6. As can be seen in Table 1 and Fig. 6, DCNN-Match Hinge and DCNN-Match CE approaches predicted a large number of landmarks per CT scan pair. In DCNN-Match Hinge + CE, the use of an auxiliary loss allows for applying an additional constraint on the landmark correspondences. Consequently, the number of predicted landmark correspondences per image was fewer than with using either of the loss separately. Further, the DCNN-Match Hinge0.1 + CE and DCNN-Match Hinge0.2 + CE predicted even fewer landmarks per CT scan pair, possibly due to the additional constraint posed by the positive margin m_{pos} used in the Hinge loss. It should be noted that irrespective of the differences within different DCNN-Match variants, a considerable number of landmark correspondences were predicted by all of them in both the simulated as well as the clinical deformations test set.

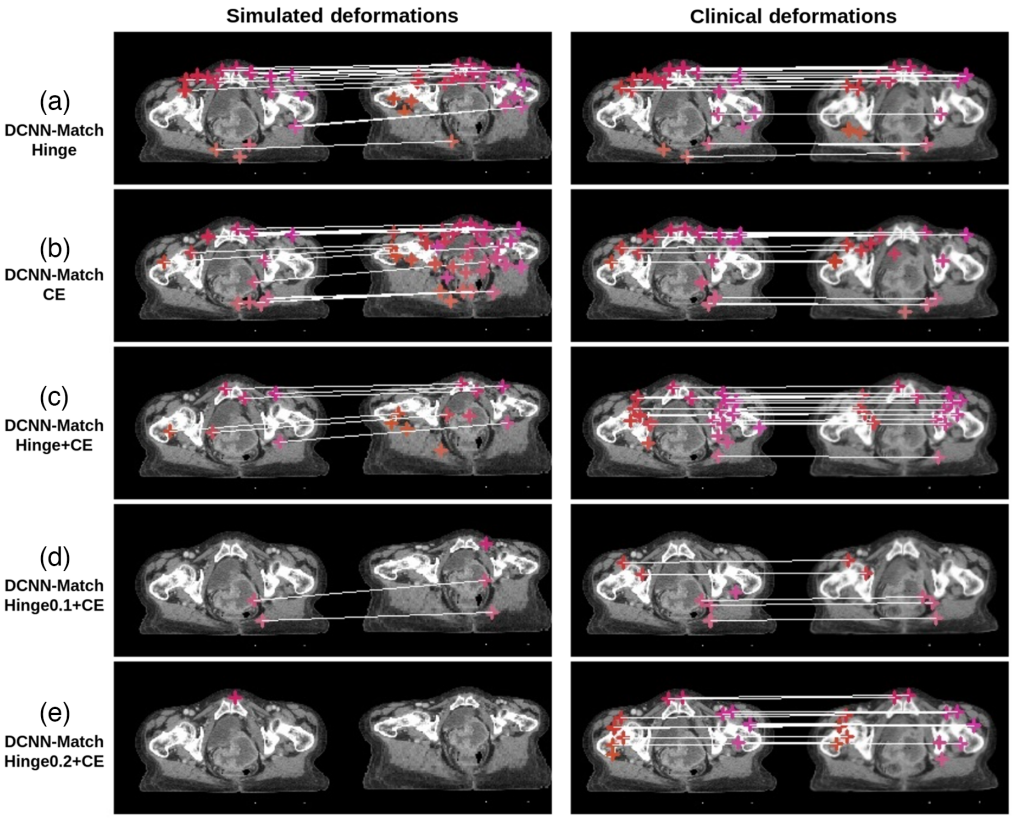

Fig. 5 Visualization of predicted landmark correspondences by (a) DCNN-Match Hinge, (b) DCNN-Match CE, (c) DCNN-Match Hinge + CE, (d) DCNN-Match Hinge0.1 + CE, and (e) DCNN-Match Hinge0.2 + CE. A transverse slice from target and source CTs in the simulated deformations test set (left) and the clinical deformations test set (right) is shown. The corresponding landmarks are shown with the same colored cross-hairs in target and source image and a white line is drawn for in-slice corresponding landmarks. It is noteworthy that some corresponding landmarks may lie on a different slice and are therefore not visible in the figure.

Table 1 Number of predicted landmark correspondences per CT scan pair. Mean (M) ± standard deviation (SD), and range (5th percentile–95th percentile) are provided.

	DCNN-Match Hinge	DCNN-Match CE	DCNN-Match Hinge + CE	DCNN-Match Hinge0.1 + CE	DCNN-Match Hinge0.2 + CE
Simulated deformations					
M ± SD	5488 ± 2258	7761 ± 2540	1698 ± 888	1735 ± 959	1220 ± 871
Range	2160–9580	2999–11400	595–3462	563–3563	244–3028
Clinical deformations					
M ± SD	3708 ± 1052	7427 ± 1682	946 ± 391	1062 ± 479	511 ± 307
Range	2563–5344	5394–10340	491–1569	455–1819	193–1000

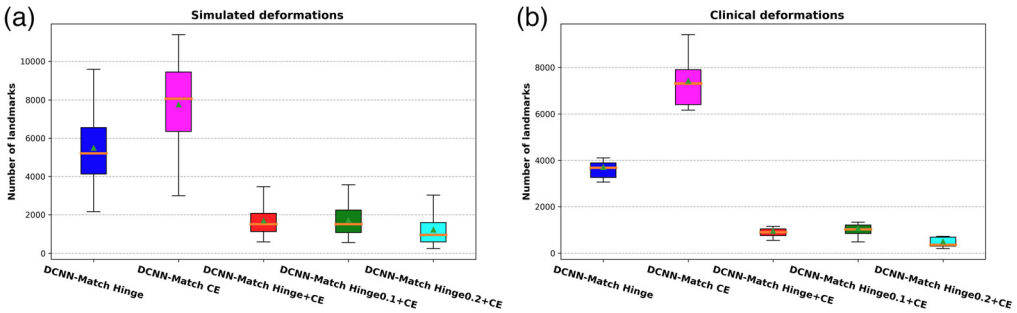

Fig. 6 Distribution of predicted automatic landmark correspondences across patients in (a) simulated deformations test set and (b) clinical deformations test set. The boxes extend from the lower to upper quartile values of the data, with a line at the median. Mean is shown with a triangular marker and whiskers represent the range from 5th percentile to 95th percentile.

3.2 *Spatial Matching Errors of Landmark Correspondences*

The cumulative distribution of the predicted landmark correspondences in the simulated test set is plotted against their spatial matching errors in Fig. 7. Both DCNN-Match Hinge and DCNN-Match CE predicted more than 70% landmarks with <2 voxels (equivalent to 4 mm) spatial matching error. But, DCNN-Match CE predicted a higher percentage of landmarks within a specific spatial matching error as compared to DCNN-Match Hinge. The decrease in spatial matching errors could be attributed to the added parameters used in the dedicated descriptor

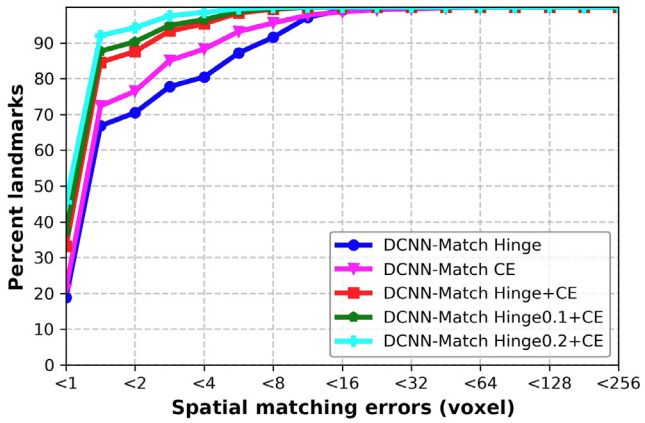

Fig. 7 Cumulative distribution of the landmarks with respect to the spatial matching errors for different versions of DCNN-Match on the simulated deformations test set-CT.

matching module in DCNN-Match CE as opposed to the parameter-free module in DCNN-Match Hinge. Further, DCNN-Match Hinge + CE takes advantage of the auxiliary loss and therefore, the landmark correspondences are predicted with lower spatial matching errors. About 90% of the predicted landmarks had a spatial matching error of <4 mm.

As expected, training with $m_{\text{pos}} > 0$ yielded landmarks with lower spatial matching errors as compared to DCNN-Match Hinge + CE (Fig. 7). Specifically for DCNN-Match Hinge0.2 + CE, more than 90% of the predicted landmark correspondences had spatial matching errors of less than 1 voxel, which is equivalent to 2 mm (image resolution). This finding indicates the potential of the automatic landmark correspondences predicted by the DCNN-Match variant for use in clinical applications.

3.3 *Spatial Matching Errors in Clinical Data*

In Fig. 5, the predicted landmark correspondences from DCNN-Match variants on a representative transverse slice from the clinical deformations test set are shown for the reader's perusal. More examples are shown in Figs. 8(a) and 8(d). The border colors indicate the ranking given by the clinician: green = good, yellow = moderate, red = wrong match. Figure 8(a) demonstrates a good match in the small bowel despite the difference of the underlying contrast and Fig. 8(b) demonstrates a good match in the muscle despite a change in the muscle deformation. In Fig. 8(c), both the landmarks are present in the rectum, but in different locations, although it was challenging to review because of the presence of the gas pocket and change in the muscle deformation. Figure 8(d) shows an example of a wrong match in the bowel. It is important to note the underlying challenges visible between the two scans in Fig. 8(d), e.g., difference in contrast, and content mismatch due to presence of gas pockets.

Figure 8(e) shows that more than 72% landmark correspondences were ranked as good match i.e., approximately within 5 mm distance and about 90% landmark correspondences were ranked

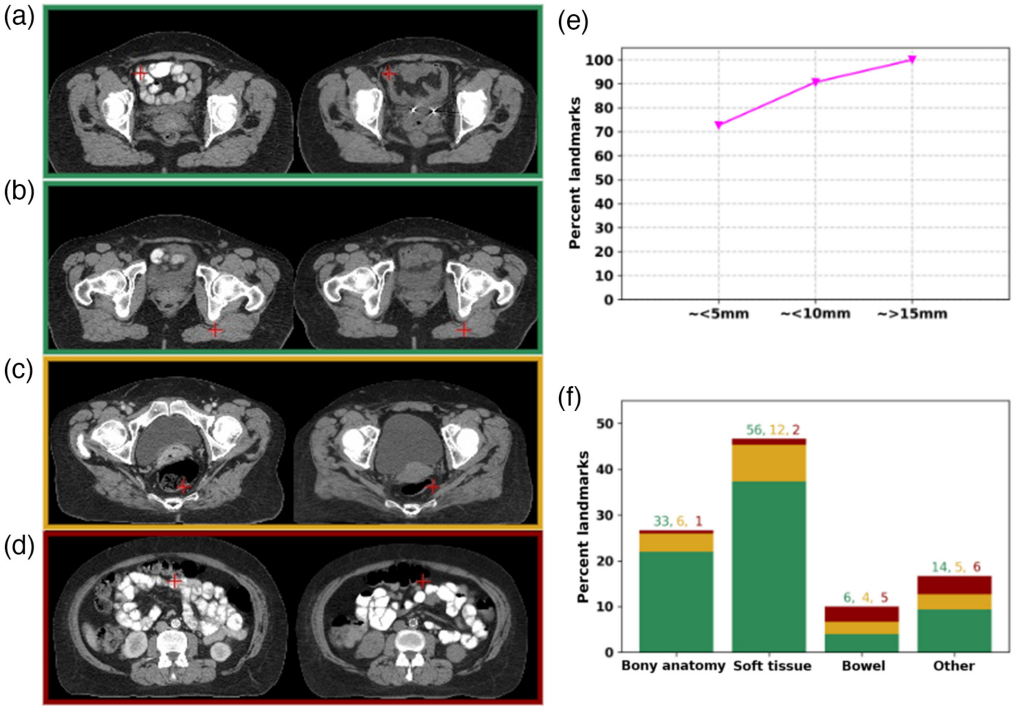

Fig. 8 Validation of landmark correspondences in clinical deformations test set. Representative examples of (a) and (b) good match, despite contrast variation and difference in muscle deformation; (c) moderate match; and (d) wrong match. (e) Cumulative distribution of landmarks with respect to (approximate) spatial matching errors. (f) Distribution of landmarks in different anatomical categories. The bars are shaded in proportion to the number of landmarks corresponding to a rank: green = good, yellow = moderate, red = wrong. In each anatomical category, the total number of landmarks representing a rank is provided in the text above bars in the corresponding color.

to be within 10 mm distance. These results indicate only a small performance difference in comparison to the simulated deformations test set [magenta curve in Figs. 7 and 8(e)], which is expected due to the presence of additional challenges in the clinical data.

Further, in Fig. 8(f), the percentage of landmarks in bony anatomy, soft tissue, bowel, and other regions is plotted. The bars in the plot are shaded in green, yellow, and red colors in proportion to the ranking of the landmarks (green = good, yellow = moderate, red = wrong) in that anatomical category. It is worth noting that the wrong matches are mainly in the bowel region, where content mismatch may happen along with large deformations and intensity variations.

3.4 *Effect of Landmark Correspondences on DIR*

In Table 2, the TRE values in the simulated and clinical deformations test sets are provided. In Fig. 9, boxplots of TRE values are provided. In both test sets, there was a significant main effect of the experimental scenario (i.e., DIR without landmarks and with landmarks predicted by either one of the DCNN-Match variants) on the observed TRE_{after} values, ($\chi(5) = 6620.117$, $p = 0e^0$) in the simulated test set and [$\chi(5) = 34.051$, p = 0.000002] in the clinical test set. It is noteworthy that in the simulated test set, the sample size was quite large (100 landmarks per scan $\times$ 121 scans = 12100) giving rise to near zero p values in the statistical testing.

In the simulated deformations test set, the post-hoc comparisons revealed that TRE_{after} values from registration using additional guidance from landmark correspondences predicted by any

Table 2 TREs in mm of pre-specified landmarks (for details refer to Sec. 2.7.2) before DIR but after affine registration (TRE_{before}) and after DIR with different approaches (TRE_{after}). Mean (M) ± standard deviation (SD), and range (5^{th} percentile–95^{th} percentile) are provided. Best TRE values are highlighted in bold.

		Simulated deformations		Clinical deformations	
		M ± SD	Range	M ± SD	Range
TRE_{before}		21.99 ± 12.67	6.00–41.76	8.50 ± 5.81	2.00–19.96
TRE_{after}	Without landmarks	5.07 ± 9.98	0.00–20.20	6.85 ± 5.79	2.00–19.12
	DCNN-Match Hinge	3.58 ± 8.80*	0.00–12.33	6.69 ± 5.84	2.00–19.53
	DCNN-Match CE	**3.14 ± 8.61***	0.00–10.77	**6.42 ± 5.79***	2.00–19.94
	DCNN-Match Hinge + CE	3.21 ± 8.63*	0.00–10.95	6.74 ± 5.77	2.00–19.47
	DCNN-Match Hinge0.1 + CE	3.18 ± 8.62*	0.00–10.77	6.79 ± 5.83	2.00–19.31
	DCNN-Match Hinge0.2 + CE	3.27 ± 8.65*	0.00–10.95	6.82 ± 5.86	2.00–19.53

*Represents significance in post-hoc comparison against TRE_{after} without landmarks.

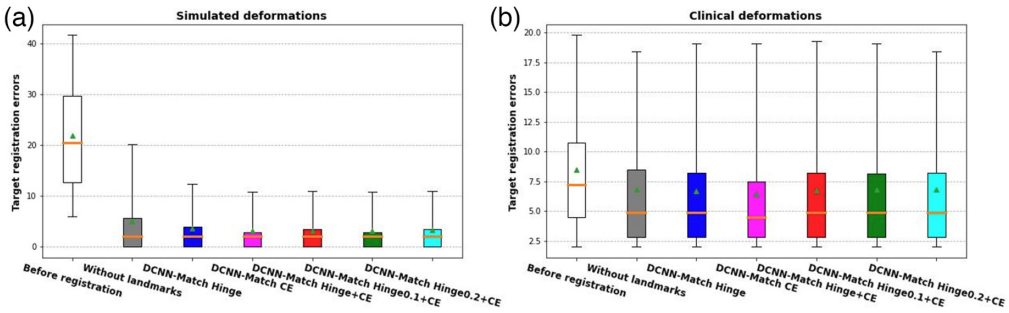

Fig. 9 Distribution of TRE in (a) simulated deformations test set and (b) clinical deformations test set. The boxes extend from the lower to upper quartile values of the data, with a line at the median. Mean is shown with a triangular marker and whiskers represent the range from 5th percentile to 95th percentile.

of the DCNN-Match variants were significantly lower than TRE_{after} values from registration without using additional guidance from landmark correspondences. However, the strongest effect was observed with landmark correspondences from DCNN-Match CE ($p = 0e^0$).

On the clinical deformations test set, although the TRE_{after} values from registration with the use of additional guidance by automatic landmark correspondences were smaller than the TRE_{after} values from registration without using additional guidance from landmark correspondences, the differences were small. Only the post-hoc comparison between TRE_{after} values from registration using landmark correspondences predicted by DCNN-Match CE and TRE_{after} values from registration without using landmark correspondences yielded statistical significance after correction for multiple comparisons ($p = 0.030$).

3.5 *Differential Effect of DCNN-Match Variants on DIR*

The post-hoc analysis indicated that the landmarks predicted by DCNN-Match CE had significantly more added value (as reflected by the TRE_{after} values) as compared to the landmarks predicted by DCNN-Match Hinge on the simulated test set ($p = 0e^0$). However, a similar finding could not be corroborated on the clinical deformations test set—pairwise comparison of TRE_{after} values obtained by DCNN-Match CE and DCNN-Match Hinge did not yield significance after correcting for multiple comparisons ($p = 0.406$).

Based on the observed spatial matching errors, it is intuitive to expect that DCNN-Match Hinge + CE would yield lower TRE values after registration as compared to DCNN-Match CE. However, surprisingly this is not the case (Table 2). TRE_{after} values using DCNN-Match CE were significantly lower than TRE_{after} values using DCNN-Match Hinge + CE in the simulated deformations test set ($p = 0.013$). In the clinical deformations test set also, the TRE_{after} values using DCNN-Match CE were significantly lower than TRE_{after} values using DCNN-Match Hinge + CE ($p = 0.046$).

Furthermore, the TRE values after registration were not affected by increasing m_{pos} in the simulated test set. The post-hoc pairwise comparisons of TRE_{after} values using DCNN-Match Hinge + CE vs DCNN-Match Hinge0.1 + CE were not significant ($p = 0.783$) on the simulated deformations test set. In fact, the TRE_{after} values using DCNN-Match Hinge0.2 + CE values were significantly higher than TRE_{after} values using DCNN-Match Hinge0.1 + CE ($p = 0.000244$). This indicates that even though an increase in m_{pos} predicts landmark correspondences with lower spatial matching errors, there is no additional benefit toward DIR performance. The observations on clinical deformations also corroborated the findings on simulated deformations. None of the post-hoc comparisons between experimental scenarios with different m_{pos} values were significantly different in the clinical deformations test set.

Overall, the results from pairwise comparisons between the TRE_{after} indicate that the added value of the automatic landmark correspondences towards the improvement of DIR performance is dependent on the underlying approach for identifying automatic landmark correspondences.

3.6 *Relation between Aspects of Automatic Landmarks and DIR Performance*

In Fig. 10(a), the landmarks correspondence accuracy (averaged over 121 patients) as described in Sec. 2.7.3 in the regions of different underlying deformation is plotted for each DCNN-Match variant. As can be seen, the correspondence accuracy of the automatic landmarks predicted by DCNN-Match Hinge deteriorated as the underlying deformation increased. A similar trend was observed for DCNN-Match CE, but to a lesser extent. As expected, the correspondence accuracy of the landmarks predicted by DCNN-Match Hinge + CE was higher than both DCNN-Match Hinge as well as DCNN-Match CE in all regions of the underlying deformation. Further, the purpose of experimenting with $m_{pos} = 0.1$ and $m_{pos} = 0.2$ to encourage high landmarks correspondence accuracy in the regions of high deformation seems to be fulfilled. The landmarks correspondence accuracy was high irrespective of the extent of the underlying deformation for DCNN-Match Hinge0.1 + CE and even higher for DCNN-Match Hinge0.2 + CE.

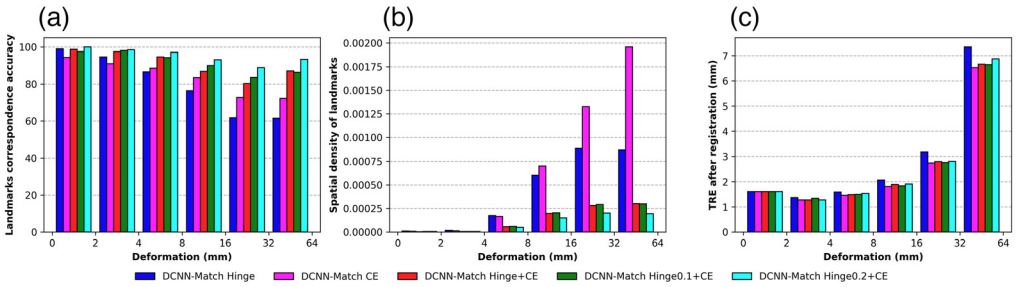

Fig. 10 Analysis of relation between aspects of landmark correspondences and DIR performance. (a) Landmarks correspondence accuracy in different regions of underlying deformation represented by the percentage of landmarks predicted within 4 mm spatial matching errors. (b) Spatial density of landmarks (number of landmarks per voxel) predicted in different regions of underlying deformation. (c) TRE of validation landmarks after DIR using automatic landmarks.

In Fig. 10(b), the spatial density of predicted landmark correspondences (averaged over 121 patients) in different regions of the underlying deformation is plotted for each DCNN-Match variant. The plot shows that DCNN-Match CE predicted more landmarks in regions with high deformations as compared to other DCNN-Match variants, which is purely empirical.

In Fig. 10(c), the TRE_{after} values of the validation landmarks (averaged over 121 patients) in different region of the underlying deformation are plotted for each DCNN-Match variant. As is apparent from the figure, the TRE_{after} values were lowest in all deformation regions when the automatic landmarks predicted by DCNN-Match CE were used as compared to the other DCNN-Match variants.

If we analyze the plots in the Fig. 10 collectively, we observe that in high deformation regions, DCNN-Match CE predicted landmarks with lower landmarks correspondence accuracy but higher spatial density as compared to DCNN-Match Hinge + CE, DCNN-Match Hinge0.1 + CE, and DCNN-Match Hinge0.2 + CE. Further, the DIR performance in the highly deformed regions was higher (reflected by lower TRE_{after} values) with the use of the automatic landmarks predicted by DCNN-Match CE as compared to DCNN-Match Hinge + CE, DCNN-Match Hinge0.1 + CE, and DCNN-Match Hinge0.2 + CE. This implies that a larger number of slightly less accurate landmarks in highly deformed regions may be more favorable for guiding the DIR approach as compared to a smaller number of highly accurate landmarks.

3.7 *Determinant of Spatial Jacobian and Qualitative Evaluation*

The determinant of the spatial Jacobian of the obtained DVFs was observed to be non-negative in all the registrations obtained in all the experimental scenarios. This indicates that all the obtained registrations were anatomically plausible.

Figure 11 shows a representative example of registration without using landmarks and registration with the DCNN-Match CE approach. The source image has a large local deformation in the center along with small random deformations globally. The transformed source images obtained after DIR have been overlaid onto the target image [columns (b) and (d)] using complementary colors such that the aligned structures look grey and misalignment is highlighted in colors. As can be seen in column (b), many regions are not aligned properly after the registration, but, with the additional guidance information [column (d)], the anatomical structures look perfectly aligned. The corresponding landmark pairs are shown with cross-hairs of the same color in the target and source images. It is worth noting that DCNN-Match CE can find landmark correspondences in highly deformed regions as well. As a result, DIR with landmark correspondences can find a better estimation of the underlying deformation field as compared to the baseline DIR approach. Columns (c) and (e) represent the root mean square errors (RMSE) of the ground truth DVF and the DVF obtained after DIR without and with landmark correspondences. Further, Fig. 12 shows an example of DIR without and with using landmarks for clinical deformations. While the output of registration without and with using landmark correspondences looks similar in most cases, a subtle improvement in alignment can still be spotted in some regions of the

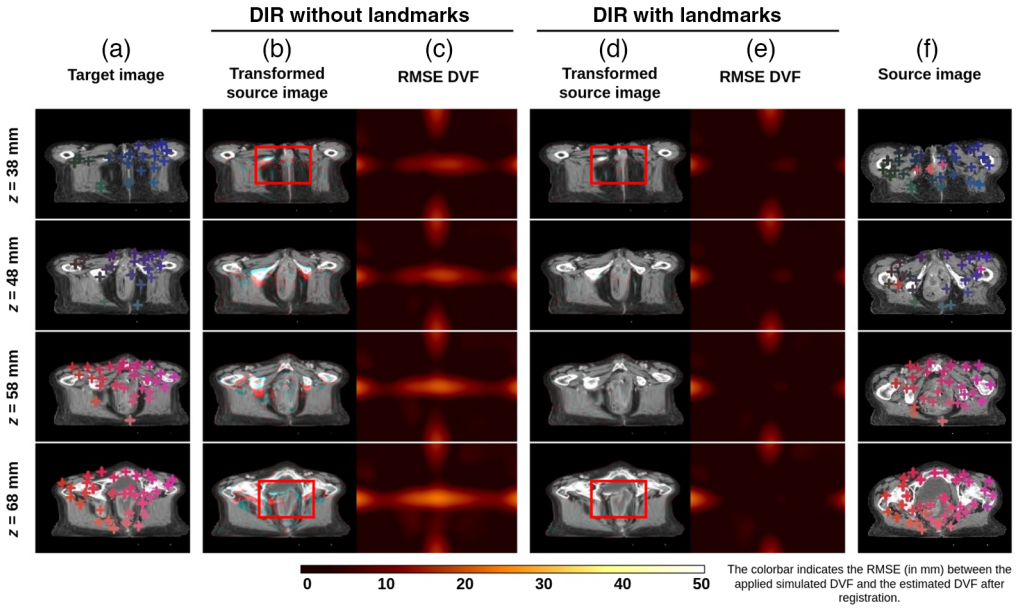

Fig. 11 Qualitative results on simulated deformations test set. Transverse slices from 10 mm apart from a representative example are shown in different rows. Column (a) target image; columns (b) and (c) RMSE plot between the ground truth and estimated DVF after registration without automatic landmarks, respectively; columns (d) and (e) transformed source image and RMSE plot between the ground truth and estimated DVF after registration with using automatic landmarks predicted by DCNN-Match CE, respectively; and column (f) source image. Landmark correspondences between the target and source images are shown in similar colored cross-hairs in columns (a) and (f). Note: some of the landmarks may have correspondences in the transverse slices not shown in the figure. The red rectangles highlight the effect of using landmark correspondences in a highly deformed region.

images (also highlighted with a red rectangle in the figure) with the use of landmark correspondences in the DIR. The determinant of the spatial Jacobian shown in Figs. 12(c) and 12(e) shows no visible image folding in the DIR solutions obtained by either of the approaches.

3.8 *Generalization to MRI Dataset*

A representative example of predicted landmark correspondences by DCNN-Match CE on MRI scans without retraining is shown in Fig. 13(a). Upon visual inspection, the predicted landmark correspondences seem to be accurate despite the different modality of the test scans. Further, the FOV of the acquisition of MRI scans was ∼16 times smaller than the FOV of the acquisition of CT scans in the test set. To make a direct comparison between the number of predicted landmark correspondences in CT and MRI datasets, we calculated the spatial density of predicted landmarks by dividing the number of landmarks by the total number of voxels in each image. In CT scan images, a large portion of the image consists of background voxels where the DCNN-Match variants do not predict landmark correspondences. Therefore, we considered only the voxels in the patient's anatomy by counting the number of voxels in the largest connected component after binarizing the image through intensity thresholding.

The spatial density of predicted landmarks in both CT and MRI test sets is shown in Fig. 13(b). Since the networks were not trained on MRI scans, the spatial density of the predicted landmarks was reduced in MRI scans. Still, a considerable number of landmarks (on average for all patients) were predicted in the MRI test set by each approach (shown as the text above bars). Further, the spatial matching errors [shown in Fig. 13(c)] of the predicted landmark correspondences on MRI scans were comparable to the spatial matching errors observed for CT scans. Overall, the above results demonstrate the generalization potential of DCNN-Match on cross-modality data without retraining.

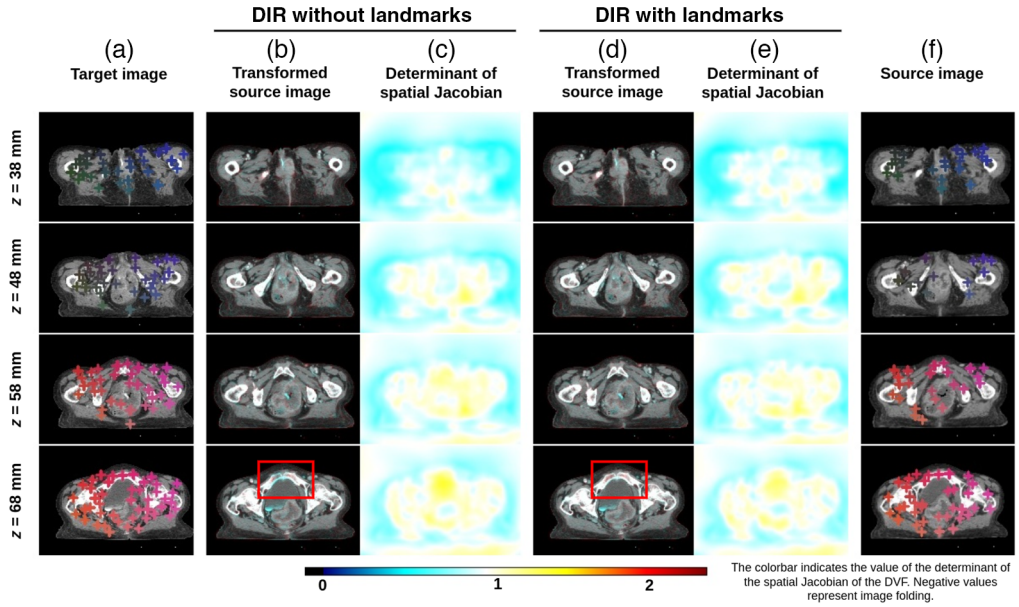

Fig. 12 Qualitative results on clinical deformations test set. Transverse slices from 10 mm apart from a representative example are shown in different rows. Column (a) target image; columns (b) and (c) transformed source image and determinant of the spatial Jacobian after registration without automatic landmarks, respectively; columns (d) and (e) transformed source image and determinant of the spatial Jacobian after registration with using automatic landmarks predicted by DCNN-Match CE, respectively; and column (f) source image. Landmark correspondences between the target and source images are shown in similar colored cross-hairs in columns (a) and (f). Note: some of the landmarks may have correspondences in the transverse slices not shown in the figure. The red rectangle highlights a region where improvement by adding landmarks correspondences in the DIR is visible.

4 Discussion

We developed a self-supervised deep-learning method (DCNN-Match) for automatic landmarks correspondence detection in 3D medical images. We have also presented quantitative and qualitative evidence that a high number of landmark correspondences with good spatial matching accuracy can be predicted within seconds with the help of our proposed approach. Furthermore, we integrated DCNN-Match with a DIR pipeline and assessed the added value of automatic landmark correspondences toward the improvement of intra-patient DIR performance. To the best of our knowledge, this is the first study to develop a self-supervised deep-learning approach for predicting automatic landmark correspondences in 3D medical images and investigating their applicability in improving DIR.

We developed five variants of the proposed approach, which differed in the way feature descriptor matching is learned. We observed that a separate module for learning feature descriptor matching (DCNN-Match CE) yields landmark correspondences with not only reduced spatial matching errors but also an increased number of matches in regions of high deformation. The results also showed that the added value to the performance of DIR was most prominent by the use of automatic landmark correspondences predicted by DCNN-Match CE. While three other variants predicted automatic landmark correspondences with better spatial matching accuracy than DCNN-Match CE, the numbers of predicted landmarks by these variants were fewer than the number of landmarks predicted by DCNN-Match, especially in regions of high deformation. This implies that the spatial density of predicted landmarks with respect to the underlying deformation plays a role in the extent of the added value provided by the automated landmark correspondences.

The results also showed that the additional guidance by automatic landmark correspondences improved the performance of DIR irrespective of the variance in the number, spatial matching errors, and spatial distribution of the automatic landmarks in both simulated as well as clinical

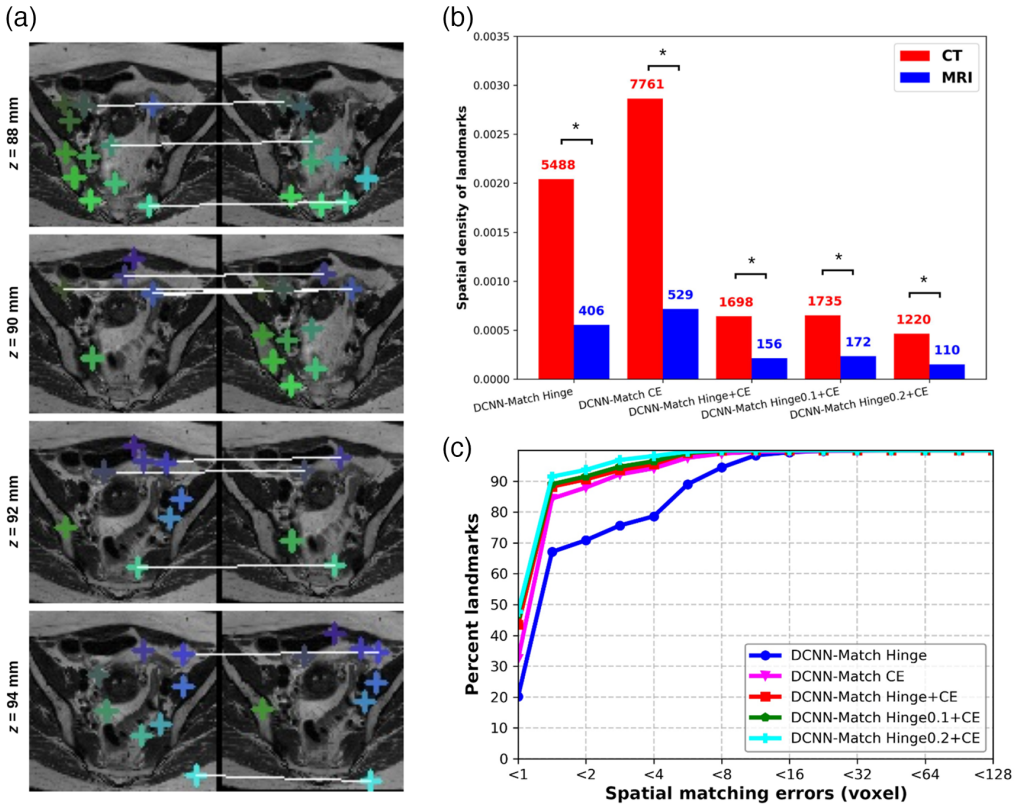

Fig. 13 Generalization results on the simulated deformations test set - MRI. (a) Predicted corresponding landmarks in the target and source MRI. Corresponding landmarks are shown with similar colored cross-hairs in the target and source images. Note that some of the landmarks match across slices following the underlying deformation in 3D. (b) Comparison of the spatial density of predicted landmarks (averaged over all patients) between simulated deformations test set – CT and simulated deformations test set – MRI for each DCNN-Match variant. The average number of predicted landmarks is shown in the text above bars. * indicates significant difference after Mann–Whitney U test. (c) Spatial matching errors of predicted landmark correspondences.

deformations test sets. These findings are in line with the existing literature on the use of automatic landmarks for the improvement of DIR in chest CT,[11,12,35] head and neck CT,[36] retinal images,[13] and brain MRI images.[14,37] A study on DIR of thoracic CT scans[38] reported that automatic landmarks-based optimization of the regularization parameter reduced the TRE of expert landmarks on average by 0.07 mm. Another study on registration of CT scans corresponding to end-inspiration and end-expiration phases reported a reduction of TRE of expert landmarks from 1.34 ± 2 mm to 0.82 ± 0.97 mm by the use of automatic landmarks in DIR.[12] Our experiments showed that the TRE of validation landmarks in the simulated deformations test set reduced from 5.07 ± 9.98 to 3.14 ± 8.61, and the TRE of expert landmarks in the clinical deformations test set reduced from 6.85 ± 5.79 to 6.42 ± 5.79 on average by the use of automatic landmark correspondences predicted by DCNN-Match CE in DIR. Since the improvement in DIR performance reported in terms of TRE values of the expert landmarks is affected by several factors, e.g., the number and location of the expert landmarks, image resolution, and TRE values before registration, a comparison in absolute values of TRE improvement cannot be made. Nevertheless, the current study adds to the existing evidence on the added value of automatic landmark correspondences in improving DIR by providing experimental results from pelvic CT scan registrations, which otherwise did not exist.

Two other studies have looked into intra-patient DIR in cervical cancer patients.[4,39] The authors in one of the studies[4] have focused on dose mapping and do not report TRE values. The average TRE values after registration reported in the other study[39] are the following: 3.5 ± 2.4 mm for bladder top, 8.5 ± 5.2 mm for cervix tip, 5.7 ± 2.1 mm for markers, and

4.6 ± 2.2 mm for the midline. As such, a direct correspondence between the landmarks used in our study and landmarks in the earlier study cannot be ascertained. Moreover, the underlying dataset and methods used are also different. Still, the mean TRE value obtained after registration with additional guidance information from landmark correspondences predicted by DCNN-MatchCE (6.42 ± 5.79 mm) seems to be within the range of reported TRE values, which gives some confidence that the obtained DIR results are satisfactory.

The extent of the added value provided by the use of automatic landmark correspondences in DIR was lower in the clinical deformations test set as compared to the simulated deformations test set. Our retrospective analysis (provided in Supplementary Material S1) revealed no obvious patterns regarding the spatial distribution of the automatic landmarks in relation to manual landmarks used for TRE calculations that could explain the lower added value of using automatic landmarks in the clinical deformations test set. The DIR performance in case of clinical deformations as reflected by TRE of manually annotated landmarks is affected by several factors e.g., choice of manual landmarks, inter- and intra-observer variation in the placement of manual landmarks, hyperparameters in the parameter map used for Elastix, limitations of Elastix in modeling large deformations, sliding tissue, and singularities in DVF. Therefore, establishing a direct relationship between the quality of automatic landmark correspondences and the DIR performance is difficult. However, we can speculate on a few factors that impacted the quality of automatic landmark correspondences and hence could have impacted the added value to DIR. In the clinical test set, the CT scans were acquired with contrast administered via a rectal tube or intravenously. Consequently, one or multiple regions (e.g., vagina, bladder, bowel bag, or vascular regions) were contrast-enhanced giving rise to large differences in appearance between the CT scan pairs, which was not a part of the training for DCNN-Match. An example of appearance variation due to contrast is shown in Fig. 4(b). This appearance variance between the source and target CT scans often overlapped with the large and complex deformations in the bladder and bowel bag. This posed an additional challenge for finding landmark correspondences between scans. Although all DCNN-Match variants were still able to find landmark correspondences in these scans despite the aforementioned challenges, they failed to find correspondences in regions where appearance was strongly different due to a combination of contrast administration and underlying deformation. We expect that incorporating a model for simulating contrast differences between scans and a better (probably a bio-mechanical based) model for simulating deformations due to physical phenomena such as bladder filling would lead to the prediction of automatic landmarks in the aforementioned challenging scenario as well and yield a larger added value of using automatic landmark correspondences in DIR. We are considering pursuing this direction for a future study.

Another factor affecting the DIR performance in the clinical deformations test set is that we tuned the hyperparameters used in Elastix (weights of the objectives used in DIR, weight_1, and weight_2) based on the DIR of CT scan pairs in the validation set consisting of simulated deformations. We used these hyperparameters for all the registrations in both simulated as well as clinical deformation test sets. This does not acknowledge the fact that each DIR problem is unique and therefore, a single setting for all source and target pairs is sub-optimal. Earlier research has also pointed out the importance of tuning the weights of different objectives in the DIR separately for each image pair to achieve the best DIR performance.[38,40] We conducted retrospective experiments by changing the weights of the objectives in DIR, which revealed that weight_1, and weight_2 values corresponding to best DIR performance (quantified in terms of minimum TRE values) were indeed different for each CT scan pair in the clinical deformations test set. Unfortunately, the tuning of weight_1, and weight_2 separately for each CT scan pair in the clinical deformations test set could not be done objectively and automatically due to the unavailability of the underlying ground truth. Note that the manually annotated landmarks were used to evaluate the DIR performance and therefore using them for tuning weight_1, and weight_2 would have produced biased results. However, the purpose of this research was not to obtain the best DIR performance for each CT scan pair but to quantify the effect of additional guidance provided by the automatic landmark correspondences. Further, the added value of the additional guidance provided by the automatic landmark correspondences may be limited by erroneous matches. While the results on the simulated data indicated the benefits of more landmarks toward DIR performance, the adverse effect of erroneous matches remains unclear. It would

be interesting to investigate in a future study how much value can be gained by removing the erroneous landmark matches either using RANSAC[41] or a deep-learning approach.[42] Another interesting direction for future research can be to simultaneously learn a deep-learning model for landmark matching as well as performing DIR. Such a model can be used to investigate how many landmarks are optimal for improving the DIR. However, care needs to be taken to avoid degeneracy because landmark matching essentially is performing DIR on a sparse grid and the optimal number of landmark matches to improve DIR could quite likely be the total number of voxels in the image.

Remarkably the proposed approach for finding automatic landmark correspondences could find automatic landmark correspondences on cross-modality data without retraining. Based on this observation, we expect that with retraining (which requires minimal effort because manual annotations are not needed), the proposed approach should be able to find automatic landmark correspondences on any type of medical imaging data. Furthermore, since a considerable number of landmarks were predicted in the MRI scans with spatial matching errors comparable to the CT scans, we expect that the use of automatic landmarks should lead to performance gain in DIR on MRI scans also. With retraining on MRI scans, we expect that the added value to the DIR performance will be similar to as observed in the CT scans.

5 Conclusion

We developed a self-supervised method for automatic landmarks correspondence detection in abdominal CT scans and investigated the effect of different variants of our automatic landmarks correspondence detection approach on the performance of DIR. The obtained results provide strong evidence for the added value of using automatic landmark correspondences in providing additional guidance information to DIR. The added value of automatic landmarks in DIR is consistent across different variants of our approach and for both simulated as well as clinical deformations. Additionally, we observed that the spatial distribution of automatic landmark correspondences with respect to the underlying deformation has a considerable effect on the extent of the added value provided by landmark correspondences. A higher number of automatic landmark correspondences in highly deformed regions has more added value than more accurate but fewer landmark correspondences. Therefore, further research in the direction of developing landmark detection approaches that are aware of the underlying deformation is recommended.

In conclusion, the current study affirms the added value of using automatic landmark correspondences for solving challenging DIR problems and provides insights into what type of landmark correspondences (in terms of spatial distribution and matching errors) may be more beneficial to DIR than others.

6 Appendix A: Elastix Parameter Maps

6.1 *A.1 Affine Registration*

```
(AutomaticParameterEstimation "true")
(AutomaticTransformInitialization "true")
(AutomaticTransformInitializationMethod "Origins")
(CheckNumberOfSamples "true")
(DefaultPixelValue 0)
(FinalBSplineInterpolationOrder 1)
(FixedImagePyramid "FixedSmoothingImagePyramid")
(ImageSampler "RandomCoordinate")
(Interpolator "LinearInterpolator")
(MaximumNumberOfIterations 1024)
(MaximumNumberOfSamplingAttempts 8)
(Metric "AdvancedMattesMutualInformation")
(MovingImagePyramid "MovingSmoothingImagePyramid")
(NewSamplesEveryIteration "true")
```

```
(NumberOfResolutions 4)
(NumberOfSamplesForExactGradient 4096)
(NumberOfSpatialSamples 4096)
(Optimizer "AdaptiveStochasticGradientDescent")
(Registration "MultiResolutionRegistration")
(ResampleInterpolator "FinalBSplineInterpolator")
(Resampler "DefaultResampler")
(Transform "AffineTransform")
```

6.2 *A.2 Deformable Image Registration*

```
(AutomaticParameterEstimation "true")
(BSplineInterpolationOrder 1)
(CheckNumberOfSamples "true")
(DefaultPixelValue 0)
(FinalBSplineInterpolationOrder 1)
(FinalGridSpacingInPhysicalUnits 8)
(FixedImageDimension 3)
(FixedImagePixelType "float")
(FixedImagePyramid "FixedRecursiveImagePyramid")
(HowToCombineTransforms "Compose")
(ImageSampler "RandomCoordinate")
(Interpolator "BSplineInterpolator")
(MaximumNumberOfIterations 300 600 900 1200)
(Metric "AdvancedMattesMutualInformation"
  "TransformBendingEnergyPenalty"
  "CorrespondingPointsEuclideanDistanceMetric")
(Metric0Weight 1)
(Metric1Weight 1)
(Metric2Weight 0.01)
(MovingImageDimension 3)
(MovingImagePixelType "float")
(MovingImagePyramid "MovingRecursiveImagePyramid")
(NewSamplesEveryIteration "true" "true" "true" "true")
(NumberOfHistogramBins 32 32 32 32)
(NumberOfResolutions 4)
(NumberOfSpatialSamples 5000 5000 5000 5000)
(Optimizer "StandardGradientDescent")
(Registration "MultiMetricMultiResolutionRegistration")
(ResampleInterpolator "FinalBSplineInterpolator")
(Resampler "DefaultResampler")
(SP_A 100 200 300 400)
(SP_a 35000 30000 25000 20000)
(SP_alpha 0.602 0.602 0.602 0.602)
(ShowExactMetricValue "false" "false" "false" "false")
(Transform "BSplineTransform")
(UpsampleGridOption "true")
```

7 Appendix B: List of manually annotated landmarks

1. Fiducial markers in the vaginal wall near the cervix at the locations: posterior left, anterior mid, posterior right, posterior mid, anterior left, and anterior right,
2. bifurcation aorta,
3. os coccygis,
4. medial tip of right and left trochanter minor,

5. most caudal, dorsal, and ventral part of the corpus of lumbar vertebrae 3,
6. most caudal, dorsal, and ventral part of the corpus of lumbar vertebrae 5,
7. right and left bifurcation vena iliaca communis,
8. right and left bifurcation of artery iliaca communis,
9. umbilicus,
10. caudal tip of right and left kidney,
11. external and internal anal sphincter,
12. cervical ostium,
13. external and internal urethral ostium,
14. right and left ureteral ostium, and
15. uterus top.

Disclosures

The authors have no relevant financial interests in the manuscript. The authors have no conflict of interests.

Acknowledgments

The research is part of the research programme, Open Technology Programme with project number 15586, which is financed by the Dutch Research Council (NWO), Elekta, and Xomnia. Further, the work is co-funded by the public-private partnership allowance for top consortia for knowledge and innovation (TKIs) from the Ministry of Economic Affairs. This work was presented (in part) at the Conference of SPIE Medical Imaging 1131303: Image-Guided Procedures, Robotic Interventions, and Modeling (February 15 to February 20, 2020, Houston, Texas, USA). This work was supported by Elekta (Elekta AB, Stockholm, Sweden) and Xomnia (Xomnia B.V., Amsterdam, the Netherlands). Elekta and Xomnia were not involved in the study design, data collection, analysis and interpretation, and writing of this article.

Code, Data, and Materials Availability

The code for DCNN-Match and the 3D version is available on GitHub at: https://github.com/monikagrewal/End2EndLandmarks.

References

1. M. Chao, Y. Xie, and L. Xing, "Auto-propagation of contours for adaptive prostate radiation therapy," *Phys. Med. Biol.* **53**(17), 4533 (2008).
2. S. Ghose et al., "A review of segmentation and deformable registration methods applied to adaptive cervical cancer radiation therapy treatment planning," *Artif. Intell. Med.* **64**(2), 75–87 (2015).
3. M. Thor et al., "Evaluation of an application for intensity-based deformable image registration and dose accumulation in radiotherapy," *Acta Oncol.* **53**(10), 1329–1336 (2014).
4. B. Rigaud et al., "Deformable image registration for dose mapping between external beam radiotherapy and brachytherapy images of cervical cancer," *Phys. Med. Biol.* **64**(11), 115023 (2019).
5. I. J. Chetty and M. Rosu-Bubulac, "Deformable registration for dose accumulation," *in Semin. Radiat. Oncol.* **29**(3), 198–208 (2019).
6. S. Klein et al., "Elastix: a toolbox for intensity-based medical image registration," *IEEE Trans. Med. Imaging* **29**, 196–205 (2010).
7. T. Vercauteren et al., "Diffeomorphic demons: efficient non-parametric image registration," *NeuroImage* **45**(1), S61–S72 (2009).

8. O. Weistrand and S. Svensson, "The ANACONDA algorithm for deformable image registration in radiotherapy," *Med. Phys.* **42**(1), 40–53 (2015).
9. T. Alderliesten, P. A. N. Bosman, and A. Bel, "Getting the most out of additional guidance information in deformable image registration by leveraging multi-objective optimization," *Proc. SPIE* **9413**, 94131R (2015).
10. R. Werner et al., "Assessing accuracy of non-linear registration in 4D image data using automatically detected landmark correspondences," *Proc. SPIE* **8669**, 86690Z (2013).
11. T. Polzin et al., "Combining automatic landmark detection and variational methods for lung CT registration," in *Fifth Int. Workshop on Pulmonary Image Anal.*, pp. 85–96 (2013).
12. J. Rühaak et al., "Estimation of large motion in lung CT by integrating regularized keypoint correspondences into dense deformable registration," *IEEE Trans. Med. Imaging* **36**(8), 1746–1757 (2017).
13. Á. S. Hervella et al., "Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement," *Procedia Comput. Sci.* **126**, 97–104 (2018).
14. D. Han et al., "Robust anatomical landmark detection with application to MR brain image registration," *Comput. Med. Imaging Graphics* **46**, 277–290 (2015).
15. D. Yang et al., "A method to detect landmark pairs accurately between intra-patient volumetric medical images," *Med. Phys.* **44**(11), 5859–5872 (2017).
16. B. Bier et al., "X-ray-transform invariant anatomical landmark detection for pelvic trauma surgery," in *21st Int. Conf. Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham, pp. 55–63 (2018).
17. V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA* **316**(22), 2402–2410 (2016).
18. A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature* **542**(7639), 115 (2017).
19. A. Tuysuzoglu et al., "Deep adversarial context-aware landmark detection for ultrasound imaging," *Lect. Notes Comput. Sci.* **11073**, 151–158 (2018).
20. F. C. Ghesu et al., "An artificial agent for anatomical landmark detection in medical images," *Lect. Notes Comput. Sci.* **9902**, 229–237 (2016).
21. M. Grewal et al., "An end-to-end deep learning approach for landmark detection and matching in medical images," *Proc. SPIE* **11313**, 1131328 (2020).
22. K. Yan et al., "SAM: Self-supervised learning of pixel-wise anatomical embeddings in radiological images," *IEEE Trans. Med. Imaging* **41**(10), 2658–2669 (2020).
23. J. Devine et al., "A registration and deep learning approach to automated landmark detection for geometric morphometrics," *Evol. Biol.* **47**(3), 246–259 (2020).
24. R. Bhalodia et al., "Leveraging unsupervised image registration for discovery of landmark shape descriptor," *Med. Image Anal.* **73**, 102157 (2021).
25. D. P. Shamonin et al., "Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease," *Front. Neuroinf.* **7**, 50 (2014).
26. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
27. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision* **60**(2), 91–110 (2004).
28. K. Marstal et al., "SimpleElastix: a user-friendly, multi-lingual library for medical image registration," in *IEEE Conf. Comput. Vision and Pattern Recognit. Workshops*, pp. 574–582 (2016).
29. P. Thevenaz and M. Unser, "Optimization of mutual information for multiresolution image registration," *IEEE Trans. Image Process.* **9**(12), –2083 (2000).
30. A. Paszke et al., "Automatic differentiation in PyTorch," in *Adv. Neural Inf. Process. Syst.-W* (2017).
31. K. He et al., "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 1026–1034 (2015).
32. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Int. Conf. Learn. Represent.* (2015).
33. IBM Corp., "IBM SPSS statistics," https://www.ibm.com/nl-en/analytics/spss-statistics-software (2020).

34. O. J. Dunn, "Multiple comparisons using rank sums," *Technometrics* **6**(3), 241–252 (1964).
35. J. Ehrhardt et al., "Automatic landmark detection and non-linear landmark- and surface-based registration of lung CT images," in *Med. Image Anal. for the Clin.-A Grand Challenge, MICCAI 2010*, pp. 165–174 (2010).
36. V. Kearney et al., "Automated landmark-guided deformable image registration," *Phys. Med. Biol.* **60**(1), 101 (2014).
37. D. Han et al., "Robust anatomical landmark detection for MR brain image registration," *Lect. Notes Comput. Sci.* **8673**, 186–193 (2014).
38. A. Schmidt-Richberg et al., "Landmark-driven parameter optimization for non-linear image registration," *Proc. SPIE* **7962**, 79620T (2011).
39. L. Bondar et al., "A symmetric nonrigid registration method to handle large organ deformations in cervical cancer patients," *Med. Phys.* **37**(7 Part1), 3760–3772 (2010).
40. K. Pirpinia et al., "The feasibility of manual parameter tuning for deformable breast MR image registration from a multi-objective optimization perspective," *Phys. Med. Biol.* **62**(14), 5723 (2017).
41. M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM* **24**(6), 381–395 (1981).
42. K. M. Yi et al., "Learning to find good correspondences," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2666–2674 (2018).

Monika Grewal is a PhD student in the Evolutionary Intelligence research group at Centrum Wiskunde & Informatica, Amsterdam, The Netherlands. She received her B.Tech and M.Tech degrees in electronics and communication engineering in 2010 and 2012, respectively. The focus of her PhD research is to develop DIR methods for advancing and simplifying the radiotherapy treatment workflow, especially for cervical cancer patients. Her research interests include artificial intelligence (specifically deep learning and evolutionary algorithms) and medical imaging.

Jan Wiersma: Biography is not available.

Henrike Westerveld is a radiation oncologist at the Erasmus Medical Center, Rotterdam, the Netherlands. She received her MD from the University of Amsterdam in 2001 and her PhD in 2008. She has worked at the AKH in Vienna (A), the UMCU in Utrecht (NL) and the AMC (now Amsterdam UMC) (NL). He is specialized in the treatment of gynecological and urological tumors. Her main interest is in image-guided radiotherapy and late effects in gynecological cancers.

Peter A. N. Bosman is the group leader of the Evolutionary Intelligence research group at the Dutch National Research Institute for Mathematics and Computer Science (Centrum Wiskunde and Informatica) and a professor of evolutionary algorithms at Delft University of Technology. His research concerns the design of scalable model-based evolutionary algorithms and their application, primarily in the life sciences and health domain. He has (co-)authored more than 200 refereed publications, out of which eight received best paper awards.

Tanja Alderliesten is an associate professor in the Department of Radiation Oncology, Leiden University Medical Center, the Netherlands, where she is the group leader of the Artificial Intelligence based Innovations research group. The focus of her research is translational in nature and primarily concerns the development of state-of-the-art methods and techniques from the fields of mathematics and computer science (including image processing, biomechanical modeling, optimization, and (explainable) artificial intelligence) for (radiation) oncology.