# Deep-learning-based CT-CBCT image registration for adaptive radio therapy

Kuckertz, Sven, Papenberg, Nils, Honegger, Jonas, Morgas, Tomasz, Haas, Benjamin, et al.

Sven Kuckertz, Nils Papenberg, Jonas Honegger, Tomasz Morgas, Benjamin Haas, Stefan Heldmann, "Deep-learning-based CT-CBCT image registration for adaptive radio therapy," Proc. SPIE 11313, Medical Imaging 2020: Image Processing, 113130Q (10 March 2020); doi: 10.1117/12.2549531

Event: SPIE Medical Imaging, 2020, Houston, Texas, United States

# Deep learning based CT-CBCT image registration for adaptive radio therapy

Sven Kuckertz[a], Nils Papenberg[a], Jonas Honegger[b], Tomasz Morgas[b], Benjamin Haas[b], and Stefan Heldmann[a]

[a]Fraunhofer Institute for Digital Medicine MEVIS, Lübeck, Germany
[b]Varian Medical Systems, Baden-Dättwil, Switzerland

## ABSTRACT

While deep learning based methods for medical deformable image registration have recently shown significant advances in both speed and accuracy, methods for use in radio therapy are still rarely proposed due to several challenges such as low contrast and artifacts in cone beam CT (CBCT) images or extreme deformations. The aim of image registration in radio therapy is to align a baseline CT and low-dose CBCT images, which allows contours to be propagated and applied doses to be tracked over time. To this end, we present a novel deep learning method for multi-modal deformable CT-CBCT registration. We train a CNN in weakly supervised manner, aiming to optimize an edge-based image similarity and a deformation regularizer including a penalty for local changes of topology and foldings. Additionally, we measure the alignment of given segmentations, facing the problem of extreme deformations. Our method receives only CT and a CBCT images as input and uses ground-truth segmentations exclusively during training. Furthermore, our method is not dependent on the availability of difficult to access ground-truth deformation vector fields. We train and evaluate our method on follow-up image pairs of the pelvis and compare our results to conventional iterative registration algorithms. Our experiments show that the registration accuracy of our deep learning based approach is superior to iterative registration without additional guidance by segmentations and nearly as good as iterative structure guided registration that requires ground-truth segmentations. Furthermore, our deep learning based method runs approximately 100 times faster than the iterative methods.

**Keywords:** deformable image registration, deep learning, convolutional neural networks, multi-modality, radio therapy, dose accumulation

## 1. INTRODUCTION

Deformable image registration (DIR) is an important tool in radio therapy for cancer treatment. It is used for the alignment of a baseline CT scan and daily low-radiation cone beam CTs (CBCTs) which allows propagation of an irradiation plan and contours of anatomical structures, respectively. This enables tracking of applied doses over time and checking of compliance with thresholds for radiation of targets and organs at risk which is also referred to as dose accumulation.[1] Precise and efficient registration algorithms allow to overcome challenging and timely segmentation of structures in low-dose CBCT images at each fraction, accelerating and facilitating the workflow of radio therapy.

DIR in radio therapy is not an easy task bearing multiple challenges. Among those, multi-modal CT-CBCT registration goes along with low contrast and artifacts in CBCT images making a precise and meaningful measurement of image similarity difficult. Furthermore, flexible organs such as bladder can cause extreme deformations. Despite these challenges, image registration has become a method of choice for image-guided radio therapy and adaptive treatment planning over the last decades.[2] Recently, novel deep learning based methods have been proposed[3] showing potential of being superior to state-of-the-art iterative algorithms. However, in the field of registration in radio therapy very little work on deep learning based approaches has been done. Due to lack of ground-truth deformations in image registration, mostly unsupervised learning methods have been proposed where a deep network is trained by minimizing a loss function that is inspired by the objective function from iterative registration methods.[4,5] To include additional available information such as segmentation masks, so-called weakly supervised methods have been proposed and showed improved registration accuracy.[6,7]

Medical Imaging 2020: Image Processing, edited by Ivana Išgum, Bennett A. Landman,
Proc. of SPIE Vol. 11313, 113130Q · © 2020 SPIE · CCC
code: 1605-7422/20/$21 · doi: 10.1117/12.2549531

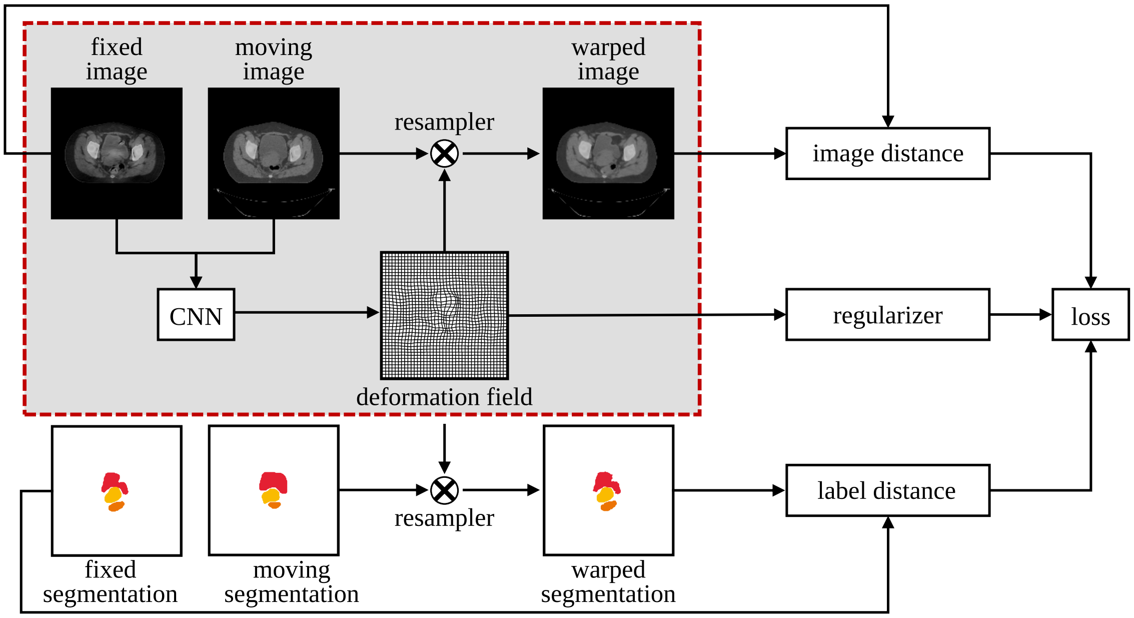


Figure 1. Schematic diagram of our weakly supervised training and inference process. A CNN is trained to minimize the loss, which is composed of an image and a label distance and a regularizer. As solely two images are forwarded to the CNN, only the subnetwork indicated by the red-dashed box is required for application of the trained CNN. Especially, no segmentations are required during inference.

Moreover, deep learning is used to overcome multi-modality. Here, deep learning is used to estimate synthetic CT images from other modalities such as MRI which then are used for registration.[8] Recently, Elmahdy *et al.*[9] proposed a patch-based deep learning method for mono-modal CT-CT image registration using a generative adversarial network and the intensity-based normalized cross correlation image similarity.

In this work we present a novel weakly supervised deep learning based method for multi-modal deformable registration of 3D CT and CBCT images. We train a CNN in weakly supervised manner based on an image similarity, deformation regularity and in addition the alignment of given ground-truth segmentations. For sake of clarity, we emphasize that the CNN receives only a CT and a CBCT image as input and the ground-truth segmentations are exclusively used for training. Our method is weakly supervised and does not require any ground-truth deformation vector fields. Our loss is composed of the normalized gradient fields (NGF)[10] image similarity measure suitable for multi-modal CT-CBCT alignment and a second term rating the alignment of given segmentation masks. Furthermore, we penalize deformation Jacobians to avoid local changes of topology and foldings. We evaluate our method on follow-up image pairs of the female pelvis and compare our results to conventional iterative registration algorithms. Our experiments show that registration accuracy of the deep learning based approach is superior to iterative registration without additional guidance by segmentations and nearly as good as iterative structure guided registration that requires ground-truth segmentations. Furthermore, the deep learning based method runs approximately 100 times faster than the iterative methods.

## 2. METHOD

Let $\mathcal{F}, \mathcal{M} : \mathbb{R}^3 \to \mathbb{R}$ denote two images, the fixed CBCT and the moving CT, respectively, and let $\Omega \subseteq \mathbb{R}^3$ be a domain modelling the field of view of $\mathcal{F}$. The aim of DIR is the generation of dense correspondences between the two images, i.e. the estimation of a deformation vector field $y : \Omega \to \mathbb{R}^3$, such that the warped moving image $\mathcal{M}(y)$ and the fixed image $\mathcal{F}$ are similar. To this end, we train a convolutional neural network that predicts the deformation between fixed and moving images. We propose a U-Net[11] based network with learnable parameters $\theta$ that takes $\mathcal{F}$ and $\mathcal{M}$ as inputs, yielding a deformation vector field $y \equiv y_\theta(\mathcal{F}, \mathcal{M})$ as output. The deformation is applied to $\mathcal{M}$ by a resampler module that differentiably warps the image using trilinear

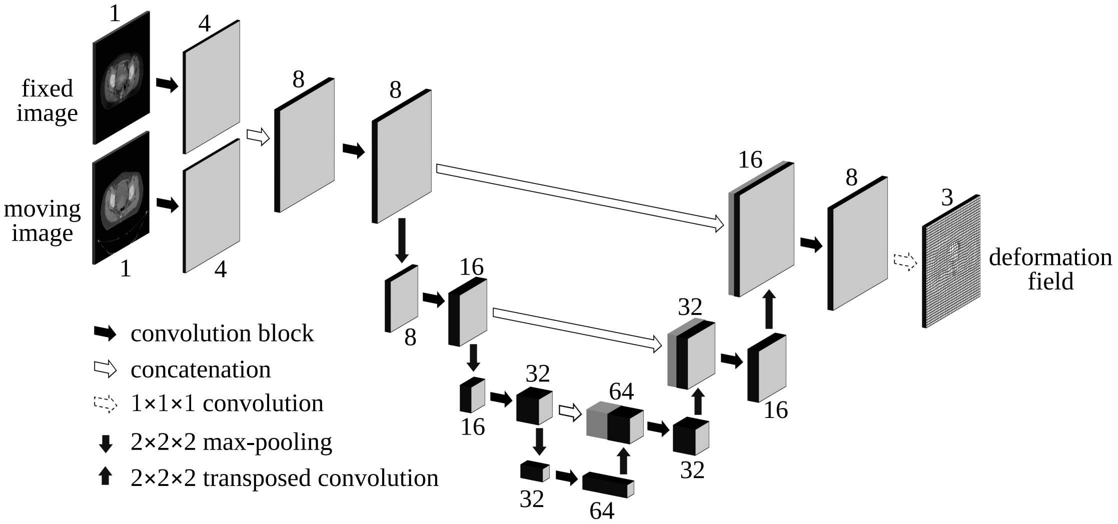


Figure 2. Proposed CNN architecture based on a U-Net.[11] Inputs are two 3D images that we want to register by the resulting deformation field. The cuboids represent multi-channel feature maps. The numbers above or below these denote the corresponding number of channels. With the depth of the network the number of channels increases, while the size of the feature maps decreases. Filled right arrows denote convolution blocks that consist of two convolutions with a kernel size of $3 \times 3 \times 3$, each followed by a ReLU and a batch normalization layer.

interpolation. Corresponding segmentations can be warped comparably using nearest-neighbor interpolation. An overview of our proposed method is shown in Fig. 1.

During training, the network parameters $\theta$ are adapted for minimizing our loss function inspired from state-of-the-art iterative image registration.[12] After training, the learned parameters $\theta$ are fixed and registration is done by a single pass of the (unseen) fixed and moving input image pair through the network. Our loss function is composed from four parts and defined as

$$\mathcal{L}(\mathcal{F}, \mathcal{M}, \mathcal{F}_{\text{seg}}, \mathcal{M}_{\text{seg}}, y) \;=\; \alpha \mathcal{D}_{\text{image}}(\mathcal{F}, \mathcal{M}(y)) \;+\; \beta \mathcal{D}_{\text{seg}}(\mathcal{F}_{\text{seg}}, \mathcal{M}_{\text{seg}}(y)) \;+\; \gamma \mathcal{R}_{\text{curv}}(y) \;+\; \delta \mathcal{R}_{\text{vol}}(y), \tag{1}$$

where $\alpha, \beta, \gamma, \delta \geq 0$ are (fixed) parameters that weight the influence of the different building blocks against each other. Due to the multi-modality of our registration problem we use the normalized gradient fields[10] image similarity measure

$$\mathcal{D}_{\text{image}}(\mathcal{F}, \mathcal{M}(y)) = \frac{1}{2} \int_{\Omega} 1 - \frac{\langle \nabla \mathcal{F}(x), \nabla \mathcal{M}(y(x)) \rangle^2_{\varepsilon_1, \varepsilon_2}}{\|\nabla \mathcal{F}(x)\|^2_{\varepsilon_1} \|\nabla \mathcal{M}(y(x))\|^2_{\varepsilon_2}} \, \mathrm{d}x, \tag{2}$$

with $\langle f, g \rangle_{\varepsilon_1, \varepsilon_2} := \sum_{j=1}^{3} f_j g_j + \varepsilon_1 \varepsilon_2$, $\|f\|_{\varepsilon_i} := \sqrt{\langle f, f \rangle_{\varepsilon_i, \varepsilon_i}}$, $i = 1, 2$ and so-called modality specific edge parameters $\varepsilon_1, \varepsilon_2 > 0$. It forces the alignment of edges in the fixed CBCT $\mathcal{F}$ and the warped CT $\mathcal{M}(y)$ and is therefore capable of registration of images with different modalities.

To achieve the alignment of important labeled structures $\mathcal{F}_{\text{seg}}$ and $\mathcal{M}_{\text{seg}}(y)$ in the fixed and warped moving image, respectively, we measure the segmentation distance $\mathcal{D}_{\text{seg}}(\mathcal{F}_{\text{seg}}, \mathcal{M}_{\text{seg}}(y))$ with a sum of squared differences of the segmentation masks that are handled as multi-channel images with one-hot-representations of the structures, i.e.

$$\mathcal{D}_{\text{seg}}(\mathcal{F}_{\text{seg}}, \mathcal{M}_{\text{seg}}(y)) = \frac{1}{2} \int_{\Omega} \|\mathcal{F}_{\text{seg}}(x) - \mathcal{M}_{\text{seg}}(y(x)))\|^2 \, \mathrm{d}x. \tag{3}$$

In order to tackle the ill-posedness of DIR and to obtain smooth deformation fields we include a curvature regularizer[13]

$$\mathcal{R}_{\text{curv}}(y) = \frac{1}{2} \int_{\Omega} \sum_{j=1}^{3} \|\Delta y_j\|^2 \, \mathrm{d}x, \tag{4}$$

Table 1. The influences of the various components within the loss function $\mathcal{L} = \alpha\mathcal{D}_{\text{image}} + \beta\mathcal{D}_{\text{seg}} + \gamma\mathcal{R}_{\text{curv}} + \delta\mathcal{R}_{\text{vol}}$ are shown by setting the weighting parameters to zero and fixing the others to their empirically determined optimal values. For Dice scores, average surface distance and the deformation Jacobians the mean and standard deviation for one fold of the cross-validation are depicted. Additionally, the average percentage of voxels in which foldings occur ($\det(\nabla y) \leq 0$) and the runtime are shown.

| Method | Dice score | Surface distance in mm | Foldings | Jacobians | Runtime in s |
|---|---|---|---|---|---|
| Affine preregistration | 0.65 ± 0.12 | 5.09 ± 2.34 | - | - | - |
| CNN, full loss | 0.81 ± 0.08 | 2.72 ± 1.32 | 0.08% | 0.997 ± 0.34 | 0.13 ± 0.14 |
| CNN, no $\mathcal{R}_{\text{vol}}$ ($\delta = 0$) | 0.80 ± 0.11 | 2.95 ± 1.71 | 1.36% | 0.999 ± 0.80 | 0.13 ± 0.14 |
| CNN, no $\mathcal{D}_{\text{seg}}$ ($\beta = 0$) | 0.68 ± 0.13 | 4.60 ± 2.31 | 0.01% | 0.999 ± 0.16 | 0.13 ± 0.14 |
| Iterative (without seg) | 0.73 ± 0.11 | 3.97 ± 1.98 | 0.01% | 0.998 ± 0.13 | 15.39 ± 2.54 |
| Iterative (with seg) | 0.86 ± 0.07 | 2.01 ± 1.12 | 0.01% | 0.998 ± 0.14 | 19.03 ± 4.39 |

penalizing high values in the second derivative of the deformation field. Because the regularizer does not eliminate physically improbable deformations like great volume changes or even grid foldings, we additionally incorporate a volume change control term[12] defined by

$$\mathcal{R}_{\text{vol}}(y) = \int_\Omega \psi(\det \nabla y(x))\, \mathrm{d}x, \tag{5}$$

where $\psi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a weighting function defined by

$$\psi(t) = \begin{cases} \frac{(t-1)^2}{t} & \text{if } t > 0, \\ \infty & \text{else.} \end{cases} \tag{6}$$

It penalizes deformation Jacobians that indicate high volume growth ($\det \nabla y > 1$), shrinkage ($0 < \det \nabla y < 1$) and grid foldings ($\det \nabla y \leq 0$).

## 3. EXPERIMENTS

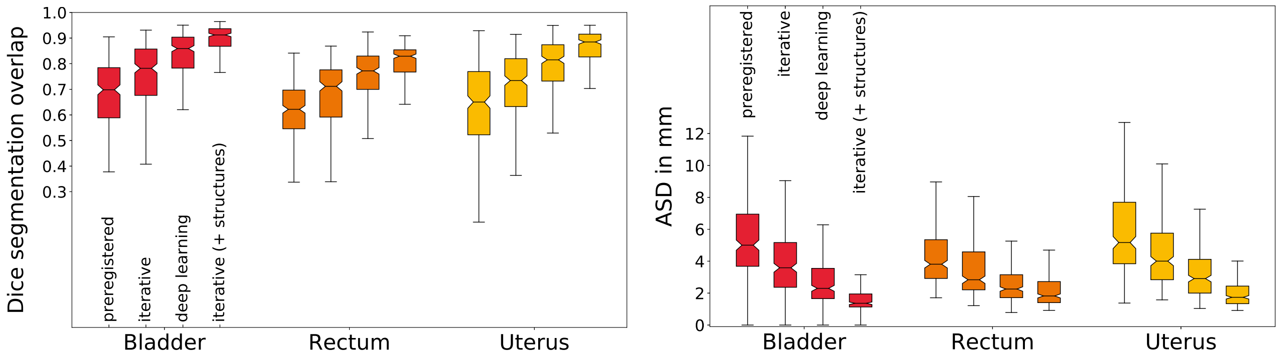
For evaluation and training, we use data from 31 female patients, acquired at multiple clinical sites. The available data for each patient consists of a planning CT and up to 26 follow-up CBCTs, yielding 256 images pairs in total. The image pairs are affinely registered, cropped to the same field of view and resampled to a size of 160 × 160 × 80 voxels, each with a size of approximately 3 mm × 3 mm × 2 mm. Additionally, we are able to integrate segmentations of the bladder, rectum and uterus (acquired by clinical experts) into the training and the evaluation of our method. We perform a fourfold cross-validation, splitting the dataset patient-wise into 4 subsets, training on 3 of them and testing on the left out subset.

For evaluation of our method we compare it to a conventional variational DIR algorithm,[14] minimizing the same loss function without a volume change control term using an iterative L-BFGS optimizer. Because we do not input segmentation masks to our deep learning based method, we also do not incorporate this information into the conventional registration process. Additionally, we analyze the benefit of actively integrating segmentation data into the iterative pipeline.

Furthermore, we analyze the influences of the different building blocks of our loss function by setting different weighting parameters to zero. The Dice similarity coefficient and the average surface distance serve as measures for the results of our experiments. The training and evaluation is implemented in PyTorch and processed on a NVIDIA GeForce RTX 2070 with 8 GB memory and an Intel Core i7-9700K with 8 cores.

## 4. RESULTS

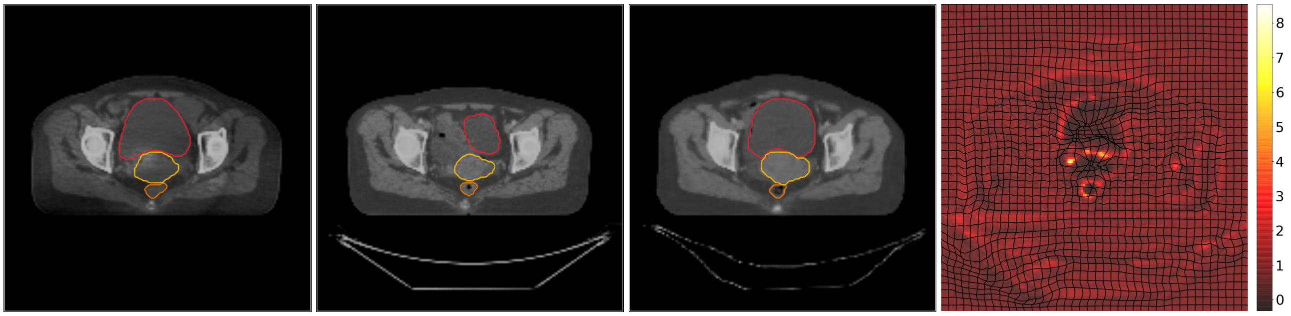
As shown in Fig. 3 our proposed deep learning based method outperforms the conventional iterative DIR algorithm both in terms of Dice overlap (in average 0.78 vs. 0.71) and average surface distance (3.10 mm vs. 4.17 mm) when there are no segmentations available for registration of the unseen test images. Here, not only



Dice Segmentation Overlap

Average Surface Distance (ASD) in mm

Figure 3. Quantitative comparison of segmentation overlap and average surface distance for all test images and annotated labels (bladder, rectum and uterus). For each one distributions after the affine preregistration, a conventional iterative and our proposed deep learning based registration are illustrated. Additionally, the results of a conventional iterative method with an active usage of segmentations are shown.



$\mathcal{F}$, $\mathcal{F}_{\text{seg}}$

$\mathcal{M}$, $\mathcal{M}_{\text{seg}}$

$\mathcal{M}(y)$, $\mathcal{M}_{\text{seg}}(y)$

$y$, $\det\nabla y$

Figure 4. Example result of our deep learning based image registration. The fixed, moving and warped images and their corresponding segmentations of the bladder 🟥, rectum 🟧 and uterus 🟨 are shown. Additionally the deformation Jacobians and the transformed grid are illustrated.

the averages are superior, also the standard deviations are smaller. Incorporating a measure for segmentation mask overlap during training enables successful alignment of organs with extreme deformations. The omission of this term results in only small improvements both in terms of Dice score and average surface distance in comparison to the affine preregistration (c.f. Tab. 1). In addition to the alignment of trained labels also other structures, e.g. body contours and bones, are well registered (c.f. Fig. 4) due to our NGF term in Eq. (1). As shown in Tab. 1, the introduction of the volume change regularizer in Eq. (1) yields smoother deformations and notably less foldings without a decrease of Dice scores or surface distances. Regarding the runtimes of the methods our framework needs in average 0.13 seconds for the registration (calculation of the deformation field and applying it) of one image pair, whereas the iterative method takes about 15 seconds.

Furthermore, Fig. 3 illustrates that the active integration of segmentation data into the conventional iterative framework yields best results referring to registration accuracy both in terms of Dice scores and average surface distance. Note that this pipeline requires knowledge about the segmentation data of both the CT and the CBCT at test time which is usually not the case in radio therapy and also no requirement for our deep learning based registration framework.

## 5. CONCLUSIONS

We presented a new deep learning based framework for 3D multi-modal DIR in radio therapy including the weakly supervised integration of additional information. Adding available segmentations into the training process showed appreciable better registration results, although only CT and CBCT images and no labeled data are needed in subsequent registration of unseen image pairs. This is a common assumption, as generating segmentations for daily CBCTs is potentially time consuming. Furthermore, we have shown that including a term for control of volume changes notably reduces the occurrence of foldings in deformation vector fields and hence increases the plausibility of generated results. Finally, including the edge-based NGF image similarity our method successfully accomplishes the challenging task of multi-modal CT-CBCT alignment for radio therapy. Our method achieves better results in comparison to state-of-the-art iterative DIR algorithms that also do not include segmentations during inference while yielding deformations over 100 times faster. However, the input of segmentations into the framework of iterative algorithms results in a notably improved registration accuracy and should therefore also be explored for our proposed method. A compromise could be the additional input of segmentations on the planning CT, since they only have to be generated once and are available for all CT-CBCT registrations afterwards. Due to the modularity of our loss function, our framework can be flexibly adapted and extended to the desired purposes which facilitates the integration of additional information. In future work we will analyze how we can use this to improve registration results in terms of both similarity and plausibility.

## REFERENCES

[1] Foskey, M., Davis, B., Goyal, L., Chang, S., Chaney, E., Strehl, N., Tomei, S., Rosenman, J., and Joshi, S., "Large deformation three-dimensional image registration in image-guided radiation therapy," *Physics in Medicine and Biology* **50**(24), 5869–5892 (2005).

[2] Brock, K. K., Mutic, S., McNutt, T. R., Li, H., and Kessler, M. L., "Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the aapm radiation therapy committee task group no. 132," *Medical physics* **44**(7), e43–e76 (2017).

[3] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A., Ciompi, F., Ghafoorian, M., Van Der Laak, J., Van Ginneken, B., and Sánchez, C., "A survey on deep learning in medical image analysis," *Medical image analysis* **42**, 60–88 (2017).

[4] de Vos, B. D., Berendsen, F. F., Viergever, M. A., Staring, M., and Išgum, I., "End-to-end unsupervised deformable image registration with a convolutional neural network," in [*Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*], 204–212, Springer (2017).

[5] Hering, A. and Heldmann, S., "Unsupervised learning for large motion thoracic ct follow-up registration," in [*Medical Imaging 2019: Image Processing*], International Society for Optics and Photonics (2019).

[6] Hering, A., Kuckertz, S., Heldmann, S., and Heinrich, M. P., "Enhancing label-driven deep deformable image registration with local distance metrics for state-of-the-art cardiac motion tracking," in [*BVM 2019*], 309–314, Springer (2019).

[7] Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., and Dalca, A. V., "Voxelmorph: a learning framework for deformable medical image registration," *IEEE transactions on medical imaging* (2019).

[8] Han, X., "Mr-based synthetic ct generation using a deep convolutional neural network method," *Medical Physics* **44**(4), 1408–1419 (2017).

[9] Elmahdy, M. S., Wolterink, J. M., Sokooti, H., Išgum, I., and Staring, M., "Adversarial optimization for joint registration and segmentation in prostate ct radiotherapy," *arXiv preprint arXiv:1906.12223* (2019).

[10] Haber, E. and Modersitzki, J., "Intensity gradient based registration and fusion of multi-modal images," in [*MICCAI 2006*], Larsen, R., Nielsen, M., and Sporring, J., eds., 726–733, Springer Berlin Heidelberg (2006).

[11] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*MICCAI 2015*], Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., eds., 234–241, Springer International Publishing (2015).

[12] Rühaak, J., Polzin, T., Heldmann, S., Simpson, I. J. A., Handels, H., Modersitzki, J., and Heinrich, M. P., "Estimation of large motion in lung ct by integrating regularized keypoint correspondences into dense deformable registration," *IEEE Transactions on Medical Imaging* **36**, 1746–1757 (Aug 2017).

[13] Modersitzki, J., [*FAIR: Flexible Algorithms for Image Registration*], SIAM (2009).

[14] König, L., Rühaak, J., Derksen, A., and Lellmann, J., "A matrix-free approach to parallel and memory-efficient deformable image registration," *SIAM Journal on Scientific Computing* **40**(3), B858–B888 (2018).