## Title

Uncertainty estimation and evaluation of deformation image registration based convolutional neural networks.



## Authors

Rivetti, Luciano
Studen, Andrej
Sharma, Manju
et al.

## Publication Date

2024-05-15

## DOI

10.1088/1361-6560/ad4c4f

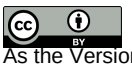
Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine



# Uncertainty estimation and evaluation of deformation image registration based convolutional neural networks

To cite this article: Luciano Rivetti *et al* 2024 *Phys. Med. Biol.*



# Uncertainty estimation and evaluation of deformation image registration based convolutional neural networks

Luciano Rivetti[1], Andrej Studen[1,2], Manju Sharma[3], Jason Chan[3], and Robert Jeraj[1,2,4]

[1]Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia
[2]Jožef Stefan Institute, Ljubljana, Slovenia
[3]Department of Radiation Oncology, University of California, San Francisco, California
[4]School of Medicine and Public Health, Department of Medical Physics, University of Wisconsin, Madison, WI

Corresponding author: Luciano Rivetti, Email: Luciano.Rivetti@fmf.uni-lj.si

## Abstract

***Objective***
Fast and accurate deformable image registration (DIR), including DIR uncertainty estimation, is essential for safe and reliable clinical deployment. While recent deep learning models have shown promise in predicting DIR with its uncertainty, challenges persist in proper uncertainty evaluation and hyperparameter optimization for these methods. This work aims to develop and evaluate a model that can perform fast DIR and predict its uncertainty in seconds.

***Approach***
This study introduces a novel probabilistic multi-resolution image registration model utilizing convolutional neural networks (CNNs) to estimate a multivariate normal distributed dense displacement field (DDF) in a multimodal image registration problem. To assess the quality of the DDF distribution predicted by the model, we propose a new metric based on the Kullback-Leibler divergence. The performance of our approach was evaluated against three other DIR algorithms (VoxelMorph, Monte Carlo Dropout, and Monte Carlo B-spline) capable of predicting uncertainty. The evaluation of the models included not only the quality of the deformation but also the reliability of the estimated uncertainty. Our application investigated the registration of a treatment planning computed tomography (CT) to follow-up cone beam CT for daily adaptive radiotherapy.

***Main results***
The hyperparameter tuning of the models showed a trade-off between the estimated uncertainty's reliability and the deformation's accuracy. In the optimal trade-off, our model excelled in contour propagation and uncertainty estimation ($p < 0.05$) compared to existing uncertainty estimation models. We obtained an average dice similarity coefficient of 0.89 and a KL-divergence of 0.15.

***Significance***
By addressing challenges in DIR uncertainty estimation and evaluation, our work showed that both the DIR and its uncertainty can be reliably predicted, paving the way for safe deployment in a clinical environment.

# 1 Introduction

Multimodal deformable image registration (DIR) is a sophisticated computational process aimed at establishing a spatial correspondence between two paired images accounting for spatial distortions. DIR is crucial for several image processing applications, such as contour propagation [1], motion tracking and modeling [2], adaptive radiotherapy [3] and lesion tracking among others. Classical registration algorithms have been extensively developed and validated to achieve high-performance registrations. Among these algorithms, B-spline models hold a special place [4], frequently utilized to perform high-quality elastic deformations in numerous open-source DIR software tools [5], [6]. These algorithms iteratively optimize an objective function that measures the quality of spatial transformations, leading to precise alignments. However, due to the iterative nature and complexity of the optimization process, they often demand significant computational resources. This results in execution times that exceed the time available in the clinical environment. As an example, in radiotherapy, the requirement of fast DIR is indispensable to perform a daily online adaptation of the treatment plan [7], [8]. Another limitation of these algorithms is their inability to provide reliable uncertainty estimation of the deformation maps.

DIR is mainly subject to two types of uncertainties. The first type is the aleatoric uncertainty, primarily stemming from limitations in the image acquisition process, including factors like a limited resolution, a poor contrast, or a high noise. These limitations make it difficult to establish a clear correspondence between the same anatomical point in two distinct images. Reducing this uncertainty would require advancements in image acquisition methods. The second type is epistemic uncertainty, often referred to as model uncertainty, which arises from the limitations of the model to find a proper registration. Reducing this type of uncertainty entails enhancing the capabilities of the employed models. The development of DIR models that can effectively capture both types of uncertainties becomes essential for the safe use of DIR algorithms in clinical applications. For example, in adaptive radiotherapy, the estimation of DIR uncertainty plays an important role in quantitatively evaluating the dose accumulated over the treatment fractions [3]. This directly impacts the plan quality evaluation and the clinical decisions for re-adaptation. In image-guided neurosurgery, DIR uncertainty estimation is crucial to prevent overconfidence in the accuracy of preoperative image matching [9]. This can help surgeons understand the registration limitations and adapt their approach to enhance surgical precision.

Specifically, to the commonly used B-spline DIR models, early efforts to estimate uncertainty consisted of evaluating the sensitivity of an image similarity metric against randomly performed variations of the optimized B-spline coefficients [10]. Although such methods can achieve high accuracy of the B-spline models, they can only provide a rough estimation of the uncertainty in subregions of the image. Another way to estimate B-spline DIR uncertainty is by doing random Monte Carlo (MC) sampling over the hyper-parameters of the B-spline model [11]. In this way, slightly different models can be optimized, and their solutions can be used to estimate the mean and standard deviation of the likely deformation. Although this method can estimate a voxel-wise DIR uncertainty, it is very time-consuming since it requires multiple runs of the registration algorithm.

With the advent of deep learning, novel algorithms have been developed for fast DIR, demonstrating competitive performance compared to classical registration methods [7], [12], [13], [14]. At the same time, considerable advancements in deep learning methods were carried out to generate models capable of estimating uncertainty. The development of Bayesian deep learning approaches offered an intuitive way to estimate models' uncertainty [15]. Unlike classical neural networks (NNs), these architectures are parametrized by a normal distribution. However, training such a network is impractical since it presents substantial computational times for training. Recent results have shown that Monte Carlo dropout models can be used to approximate Bayesian Networks and to estimate the model uncertainty [16]. Dropout layers randomly disable nodes of the network, allowing different distributions of the same model while training. When such a model is implemented multiple times during validation, it generates likely different solutions that can be used

to calculate the dispersion of the model using the standard deviation of the results. This concept was used by Gong et al. [17] to create a model that uses convolutional neural networks (CNNs) to perform DIR and predict its uncertainty. However, it is an open question how the uncertainty estimated with these models is affected by the election of the dropout probability.

Another approach to estimating uncertainty in deep learning is using Bayesian probabilistic models. Such models describe the problem using Bayesian probabilities and estimate the posterior probability distribution of the phenomenon to model using variational inference [18]. Neural networks are used to model the parameters of the approximated probability density function. One of the main features of these methods is that they allow direct estimation of the uncertainty of the problem directly from the network. Such a method was used by Dalca et al. to create a probabilistic DIR model (VoxelMorph) that can predict the deformation and its uncertainty [19]. Their innovative work not only established a bridge between classical and learning paradigms but also presented exciting possibilities for predicting DIR uncertainty quickly. Using a similar methodology, Smolders et al. built a CNN model that can predict the DIR uncertainty of deformation maps generated using classical registration algorithms [11]. They validated their model by calculating a reliability diagram that correlates the error of the propagation of landmarks drawn by a physician and the uncertainty estimated with their model.

Despite the previous efforts, the performance of the models that can predict the DIR uncertainty remains challenging to evaluate because of the absence of gold standard deformations and dedicated metrics that directly evaluate its quality. Furthermore, the relation between the predicted deformation maps' accuracy and the uncertainty estimated reliability remains unstudied. The main goal of this work is to develop and validate a fast deep learning DIR algorithm that can accurately predict the most likely deformation field and its uncertainty in a multimodal DIR problem.

# 2 Methods

In our work, we adopted the Bayesian probabilistic approach used by Dalca et al. (VoxelMorph)[19]. Their methodology describes how to predict a distribution of voxel-wise correspondences, or dense displacement fields (DDF), between a moved image ($\boldsymbol{m}$) and a fixed image ($\boldsymbol{f}$) using variational inference and CNNs. They approximated the posterior probability density function $p(\boldsymbol{\phi}|\boldsymbol{f},\boldsymbol{m})$ with a multivariate normal distribution $N(\boldsymbol{\phi};\ \mu(\boldsymbol{\psi}), \boldsymbol{\Sigma}(\boldsymbol{\psi},\sigma))$, where $\boldsymbol{\phi}$ is a sample of the DDF, $\mu$ is the mean of the distribution, $\boldsymbol{\Sigma}$ its covariance matrix, $\sigma$ a fixed scalar and $\boldsymbol{\psi}$ the (hidden) parameters of the CNN that were optimized by minimizing:

$$\mathcal{L}(\boldsymbol{\psi};\ \boldsymbol{f},\boldsymbol{m}) = \omega\,\|f - \boldsymbol{\phi}(\boldsymbol{\psi})\circ m\|^2 + Tr\big(\lambda \boldsymbol{L}\boldsymbol{C}(\sigma)\boldsymbol{\Sigma}(\boldsymbol{\psi})\boldsymbol{C}(\sigma) - \log\boldsymbol{\Sigma}(\boldsymbol{\psi})\big) + \frac{\lambda}{2}BE\big(\mu(\boldsymbol{\psi})\big) \quad (1)$$

where $\boldsymbol{f},\boldsymbol{m} \in \mathbb{R}^{1\times N}$ with N the number of voxels in the image, $\mu(\boldsymbol{\psi}) \in \mathbb{R}^{1\times 3N}$ and $\boldsymbol{\Sigma}(\boldsymbol{\psi}) \in \mathbb{R}^{3N\times 3N}$ is a diagonal matrix, $\omega$ is a scalar hyperparameter of the problem, $\|\cdot\|^2$ is the Frobenius norm, $Tr$ is the trace function, $\lambda$ is a hyper-parameter of the problem, $\boldsymbol{L}$ is the matrix representation of the Laplacian operator, $\boldsymbol{C}(\sigma)$ is a smooth convolution matrix with a standard deviation $\sigma$ and $BE\big(\mu(\boldsymbol{\psi})\big) = \mu(\boldsymbol{\psi})^T\boldsymbol{L}\,\mu(\boldsymbol{\psi})$ is the bending energy. It is important to note that $\mu(\boldsymbol{\psi})$ and $Diag(\boldsymbol{\Sigma}(\boldsymbol{\psi}))$ are predicted from a 3D UNet [20] and that $\boldsymbol{\Sigma}(\boldsymbol{\psi},\sigma) = \boldsymbol{C}(\sigma)\boldsymbol{\Sigma}(\boldsymbol{\psi})\boldsymbol{C}(\sigma)$. To derive equation (1), the authors assumed independence between the variables $\boldsymbol{\phi}$ and $\boldsymbol{m}$ and approximated $p(\boldsymbol{\phi}|\boldsymbol{m})$ with a normal distribution $N(\boldsymbol{\phi};\ \boldsymbol{0}, \lambda^{-1}\boldsymbol{L}^{-1})$.

## 2.1 Unsupervised multimodal deformable image registration framework

Although VoxelMorph provides a mathematical framework to predict the DDF mean and standard deviation, the approach has limitations on predicting a smooth DDF with a realistic uncertainty. Furthermore, it doesn't provide an efficient way to calculate $\boldsymbol{L}\boldsymbol{C}(\sigma)\boldsymbol{\Sigma}(\boldsymbol{\psi})\boldsymbol{C}(\sigma)$ from equation 1 and the approach is derived only for the unimodal registration case. Because the probability $p(\boldsymbol{\phi}|\boldsymbol{m})$ is modeled as independent of the properties of the image $\mathbf{m}$, the VoxelMorph implementation is missing important prior information that can better guide the registration uncertainty. For these reasons, we approximated $p(\boldsymbol{\phi}|\boldsymbol{m})$ with a normal distribution that is dependent on some properties of the moved image $\boldsymbol{m}$. We assumed that given a moved image $\boldsymbol{m}$, the average of the deformation

field $\boldsymbol{\phi}$ over all the possible fixed images would be zero and that the covariance matrix would depend on some properties of the moved image:

$$p(\boldsymbol{\phi}|\boldsymbol{m}) \sim N(\boldsymbol{\phi}; \mathbf{0}, \boldsymbol{\Sigma_p}^{1/2}(\boldsymbol{m})\boldsymbol{L}^{-1}\boldsymbol{\Sigma_p}^{1/2}(\boldsymbol{m})) \quad (2)$$

where $\boldsymbol{\Sigma_p}$ is a diagonal covariance matrix that depends on the properties of the moved image $\boldsymbol{m}$. We modeled $\boldsymbol{\Sigma_p}$ considering the independent contribution of two different sources of uncertainties. The first one is related to random missalignments between the moved and fixed image and its magnitude can be represented as $\boldsymbol{\Sigma_1} = d^2\mathbf{I}$, where d is the average missalignment distance and $\mathbf{I}$ is the identity matrix. The second one is related to low-contrast regions in the moved image, and it can be modeled as $\boldsymbol{\Sigma}_2 = diag(\boldsymbol{C}(d)T(\boldsymbol{m}))$, where $\boldsymbol{C}(d)$ is a smooth convolution matrix with standard deviation $d$ and $T: \mathbb{R}^{1\times N} \rightarrow \mathbb{R}^{1\times N}$ is a function that computes the distance of every voxel of the moved image to the closest neighboring voxel that generates a difference in value above a specified threshold in Hounsfields Units (HU). In this way we calculated the prior covariance matrix as $\boldsymbol{\Sigma_p} = d^2\mathbf{I} + diag(\boldsymbol{C}(d)T(\boldsymbol{m}))$.

To perform multimodal image registration between different imaging modalities, we exchanged the mean square error term of equation 1 (image similarity term) with the Mutual Information (MI). MI was chosen for its robust performance in quantifying image similarity between two multimodal images, as it maximizes the information extracted about the moved image when considering the fixed image [21]. We also considered scaling the magnitude of the BE since its value is influenced by the number of background voxels present in the images. Including these changes in equation 1, and solving the algebra (See supplementary material) we can rewrite the equation 1 as:

$$\begin{aligned}\mathcal{L}_{us}(\boldsymbol{\psi};\, \boldsymbol{f}, \boldsymbol{m}) = \omega\,(MI(\boldsymbol{f}, \boldsymbol{\phi}(\boldsymbol{\psi}) \circ \boldsymbol{m}) + \frac{\gamma}{\omega} BE(\boldsymbol{\Sigma_p}^{1/2}\mu(\boldsymbol{\psi})) \\ +\, Tr\left(\boldsymbol{\Sigma_p}^{-1}\mathrm{diag}(\boldsymbol{K}(\sigma)\boldsymbol{\Sigma}(\boldsymbol{\psi})\mathbf{1}) - \log\boldsymbol{\Sigma}(\boldsymbol{\psi})\right)\end{aligned} \quad (3)$$

where $\gamma$ is a scalar hyperparameter of the problem, $\boldsymbol{K}(\sigma) \in \mathbb{R}^{3N\times 3N}$ is an effective smoothing convolution matrix and $\mathbf{1} \in \mathbb{R}^{3N\times 1}$ is a vector full of ones. Note that the expression $\boldsymbol{K}(\sigma)\boldsymbol{\Sigma}(\boldsymbol{\psi})\,\mathbf{1}$ is equivalent to convolving the image of $\boldsymbol{\Sigma}(\boldsymbol{\psi})$ with the kernel used to generate the matrix $\boldsymbol{K}(\sigma)$. This effective 3D convolution kernel $\boldsymbol{k}(\sigma)$ can be calculated as (detailed calculation in supplementary material):

$$\boldsymbol{k}(\sigma) = \sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\sum_{k=1}^{n_z}(\delta_{\frac{n_x}{2},i}\;\delta_{\frac{n_y}{2},j}\;\delta_{\frac{n_z}{2},k} - c_{i,j,k})\,\hat{\boldsymbol{c}}(\sigma, [\frac{n_x}{2}, \frac{n_y}{2}, \frac{n_z}{2}]) \odot \hat{\boldsymbol{c}}(\sigma, [i,j,k]) \quad (4)$$

where $n_r$ is the size of the fixed or moved image in the r axis, $\delta$ is the Kroneker delta, and $\hat{\boldsymbol{c}}(\sigma, [i,j,k])$ is a Gaussian kernel centered to the position i,j,k respectively, $c_{i,j,k}$ is the i,j,k element of the centered Gaussian kernel and $\odot$ is an element-wise product. For implementation purposes, we normalized $\boldsymbol{k}(\sigma)$ to be 1 when all its elements are summed up. The incorporation of the $\mathbf{K}$ effective smoothing convolution matrix provides our method a computationally efficient way to calculate the term $\boldsymbol{L}\boldsymbol{C}(\sigma)\boldsymbol{\Sigma}(\boldsymbol{\psi})\boldsymbol{C}(\sigma)$ from equation 1 by using convolutions that can be computed in any deep learning framework in python.

Note that the MI term of equation 3 encourages image similarity between the fixed image and one realization of the warped moved image, the bending energy term regularizes $\mu(\boldsymbol{\psi})$ and the trace term, or uncertainty term, regulates the magnitude of $\boldsymbol{\Sigma}(\boldsymbol{\psi})$. This can be seen by considering the simplified 1D problem with only one voxel. In this case, $\boldsymbol{\Sigma}(\boldsymbol{\psi})$ is a scalar variable, and the uncertainty term is $\Sigma(\boldsymbol{\psi})/\Sigma p - log\,\Sigma(\boldsymbol{\psi})$. This function is minimized only when $\Sigma(\boldsymbol{\psi}) = \Sigma_p$. If we extrapolate this result, we can think that the uncertainty term pushes the smoothed variational diagonal covariance matrix $\boldsymbol{\Sigma}(\psi)$ to be similar to $\Sigma_p$. In the limit when $\omega, \gamma \rightarrow \infty$, the unsupervised loss function collapses to the sum of an image similarity metric and a DDF regularizer as it was previously formulated in other works to predict only a DDF without its uncertainty [19], [22]. However, when $\omega$ is not so big, there is a trade-off between the image similarity term and the

uncertainty term.

## 2.2 Supervised multimodal deformable image registration framework

In many clinical applications where registration is needed, the fixed and moving images have segmentation masks of anatomical regions that can be incorporated into the training process to improve the registration's accuracy. However, the process of segmentation is subjected to uncertainty in the borders of the structures, which is mostly related to the bad definitions of the anatomical borders. For this reason, we developed a probabilistic approach that incorporates the segmentation and its uncertainty using probability maps to better guide the registration distribution.

In addition to the methodology described for the unsupervised model, we propose to maximize the log-likelihood of the posterior probability distribution $p(\boldsymbol{s}_f^i \mid \phi, \boldsymbol{s}_m^i)$, where $\boldsymbol{s}_f^i \in \mathbb{R}^{1\times N}$ is the i-th contour in the fixed image and $\boldsymbol{s}_m^i \in \mathbb{R}^{1\times N}$ is the i-th contour in the moved image. Because $\boldsymbol{s}_f^i$ and $\boldsymbol{s}_m^i$ are binary contours, i.e., their voxels values are 1 or 0, we can consider $\boldsymbol{s}_f^i$ as a realization of the probability map defined as the average of different samples of the warped contour ($PM^i = \frac{\sum_k^N \phi^k \circ \boldsymbol{s}_m^i}{N}$). Following this rationale, we can model $p(\boldsymbol{s}_f^i \mid \phi, \boldsymbol{s}_m^i)$ with an independent Bernoulli distribution:

$$p(\boldsymbol{s}_f^i \mid \phi, \boldsymbol{s}_m^i) = \prod f(\boldsymbol{s}_f^i, PM) = \prod PM^{s_f^i}(1-PM)^{1-\boldsymbol{s}_f^i} \quad (5)$$

where $f$ is the voxel-wise Bernoulli distribution, $\prod$ is an element-wise multiplication and $PM$ is the probability map of the position of the contour in the fixed anatomy. Minimizing the log likelihood of $p(\boldsymbol{s}_f^i \mid \phi, \boldsymbol{s}_m^i)$ (See supplementary material), and using equation 3, we can write the loss function for the supervised model as:

$$\mathcal{L}_s(\boldsymbol{\psi};\, \boldsymbol{f},\, s_f^i,\, \boldsymbol{m},\, s_m) = \mathcal{L}_{us}(\boldsymbol{\psi};\, \boldsymbol{f}, \boldsymbol{m}) + \alpha \sum_i^M H(p\boldsymbol{S}_f^i,\, PM^i) \quad (6)$$

where $H$ is the average of the voxel-wise cross-entropy (CE), $p\boldsymbol{S}_f^i = \sum_j^J \frac{s_f^{i,j}}{J}$ is the i-th probability map of the fixed contour obtained averaging contours of J physicians on the same organ. In this work, we will use J = 1 because we only have one segmentation per organ.

## 2.3 Neural Network

The variational $\mu(\boldsymbol{\psi})$ and the diagonal of the covariance matrix $Diag(\boldsymbol{\Sigma}(\boldsymbol{\psi}))$ were modeled using CNNs. The architecture employed in this study is an adapted version of the network developed by Hu et al. [12]. By configuring the network with an output of 6 channels (3 for $\mu$ and 3 for $\boldsymbol{\Sigma}$), it becomes adept at executing probabilistic multi-resolution image registration, effectively capturing both global and local deformations in a progressive manner from coarse to fine scales. At each level of resolution j, the network predicts a multivariate normal distribution $N(\boldsymbol{\phi}^j;\, \mu^j(\boldsymbol{\psi}), \boldsymbol{\Sigma}^j(\boldsymbol{\psi}, \sigma))$ that establishes the registration from $m$ to $f$ at that resolution. The composite distribution is derived by aggregating the contributions across various resolutions, resulting in $N(\boldsymbol{\phi};\, \mu(\boldsymbol{\psi}), \boldsymbol{\Sigma}(\boldsymbol{\psi}, \sigma)) = \sum_j N(\boldsymbol{\phi}^j;\, \mu^j(\boldsymbol{\psi}), \boldsymbol{\Sigma}^j(\boldsymbol{\psi}, \sigma)) = N(\boldsymbol{\phi};\, \sum_j \mu^j(\boldsymbol{\psi}), \sum_j \boldsymbol{\Sigma}^j(\boldsymbol{\psi}, \sigma))$ The network takes the fixed and moved images as inputs and predict $\mu(\boldsymbol{\psi})$ and $\boldsymbol{\Sigma}(\boldsymbol{\psi})$ as outputs (Figure 1). These outputs are subsequently employed to construct a differentiable dense displacement field sample ($\boldsymbol{\phi}$) using the reparameterization step: $\boldsymbol{\phi} = \mu(\boldsymbol{\psi}) + \boldsymbol{C}(\sigma)\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\psi})\boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is a sample of the standard normal distribution $N(\boldsymbol{0}, \boldsymbol{I})$ [19]. This resultant sample is then utilized to warp the moved image and structures using a resampler layer. Following this step, the loss function is computed, and $\boldsymbol{\psi}$ is updated accordingly.

## 2.4 Data

Validation studies were conducted on head and neck patients undergoing external beam radiation therapy. The anonymized data was retrospectively acquired under an approved institutional review board. The moved image and contours were the planning computed tomography (pCT) and the planning structures, respectively. Each pCT contained at least ten planning structures that were segmented by an experienced radiation oncologist. The fixed image and contours were the daily cone beam computed tomography (CBCT) and their associated structures obtained by propagating the planning contours using a certified DIR algorithm (ANACONDA) that is used in clinics to perform contour propagation [23]. Every patient in this dataset contained at least one pCT with its structures and around 30 daily CBCTs with their structures.

The networks were trained with 220 pairs of images (pCT and CBCT) obtained from 9 patients. To increase the diversification of the training pairs, the pCTs and their structures were anatomically modified by warping them to different daily anatomies using Elastix. Then, the training pairs were sampled between the different anatomical variations of the pCTs and the daily CBCTs.

For validation and testing, the sampling was different from the training. The image pairs were generated as they will be in the clinics, i.e., the pCT was registered to every daily CBCT. This resulted in an independent validation dataset of 98 pairs of images (pCT and CBCT) obtained from 3 patients and an independent test dataset with 66 pairs of images obtained from 2 patients. To guarantee the fidelity of the evaluation, the testing and validation of daily CBCT structures were reviewed and corrected by an experienced radiation oncologist.

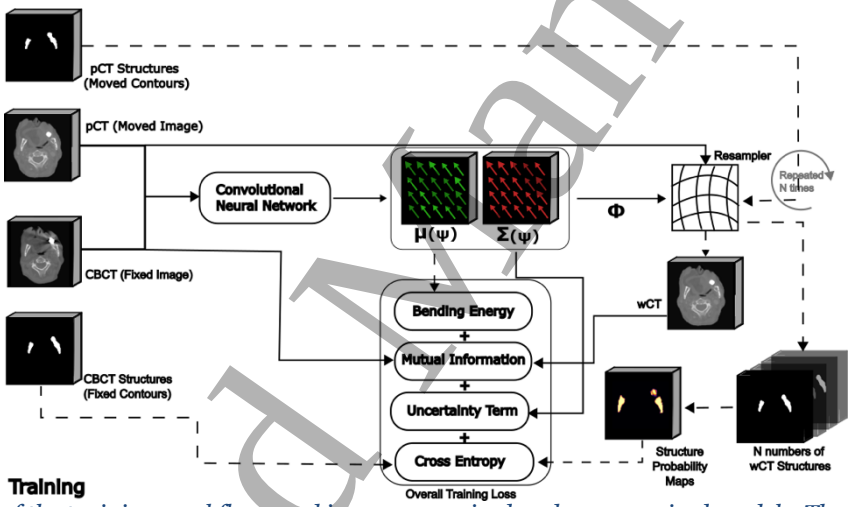



*Figure 1: Illustration of the training workflow used in our supervised and unsupervised models. The solid lines show the data flow for the unsupervised model and the solid lines plus the dash lines show the data flow for the supervised model.*

All images and structures were resampled to achieve an isotropic voxel spacing of 1.15 mm, resulting in a size of 192 voxels along each dimension. The image intensity was normalized to ensure values fell within the range of 0 to 1.

## 2.5 Validation and Testing

For validation and testing, the data flow was similar to the one described in Figure 1, with the difference that sixty samples of the pCT and its corresponding structures were generated to evaluate their dispersion. This number was selected as the maximum sampling size that our GPU could efficiently handle at once. Several metrics were used to assess the quality of the dense displacement field (DDF) distribution. The Dice Similarity Coefficient (DSC) between the propagated structures with the mean DDF ($\mu \circ \boldsymbol{s}_m^i$) and the physician's structures was calculated as an indirect measure of the quality of μ. The MI was calculated over a sample of the wCT ($\boldsymbol{\phi}(\boldsymbol{\psi}) \circ \boldsymbol{m}$), and the daily CBCTs were calculated as an indirect measure of the quality of the DDF samples. The average value of the percentage of negative values in the determinant of the Jacobian ($|J| \leq 0$) in different samples of the DDF was calculated to evaluate the smoothness of the sampled DDFs. The expected calibration error (ECE) of probability maps [24] was used as an indirect measure to assess the quality of the DDF distribution.

While it would be ideal to assess the DIR distribution based on ground truth DDFs, it is important to acknowledge that such information is largely unavailable in most datasets. Consequently, various methods have been introduced to assess DDF accuracy using DDFs derived from plausible anatomical deformations as a reference [25], [26]. However, these methods primarily evaluate the accuracy of individual realizations of the DDF rather than the entire distribution which is our aim. Additionally, they are constrained to unimodal image registration since applying the known DDF to the fixed image generates a moved image of the same modality as the fixed image. To solve these challenges, we specifically developed a probabilistic approach to enable the evaluation of DDF distribution quality in multimodal image registration problems.

The method consisted of creating a simulated phantom image ($\Omega \circ \boldsymbol{f}$) with a likely anatomical deformation ($\Omega$) obtained by registering $\boldsymbol{f}$ to a follow-up image of the same patient (Figure 2). Then, a DIR model was used to estimate the DDF distribution which registers $\boldsymbol{f}$ to $\boldsymbol{m}$ and $\Omega \circ \boldsymbol{f}$ to $\boldsymbol{m}$ ($\boldsymbol{\phi_1}$ and $\boldsymbol{\phi_2}$ respectively). Considering that $\boldsymbol{\phi_1}$ and $\boldsymbol{\phi_2}$ are diffeomorphism, then the deep learning DDF distribution ($\Omega^*$) which registers $\boldsymbol{f}$ with $\Omega \circ \boldsymbol{f}$ as $\Omega^* = \boldsymbol{\phi_2} - \boldsymbol{\phi_1}$ can be estimated. If $\Omega^*$ is a good estimation of $\Omega$, then $\mu_{\Omega^*} \approx \Omega$ and $z(\Omega^*) \sim N(\boldsymbol{0}, \boldsymbol{I})$ where $z(\Omega^*) = \frac{\boldsymbol{\Omega}^*_{sampled} - \boldsymbol{\Omega}}{\sqrt{Diag(\boldsymbol{\Sigma}_{\Omega^*})}}$ and $Diag$ is a function which set all off-diagonal elements to zero (detailed calculation in the supplementary material). In other words, the distribution generated with the elements of the vector $z(\Omega^*)$ should follow a normal distribution $N(0,1)$. For this reason, the Kullback–Leibler divergence $KL(z(\Omega^*) \parallel N(0,1))$ was used to evaluate the quality of a multivariate normal DDF distribution. In this work, this metric was averaged over all the likely anatomical deformations ($\Omega$) between the first CBCT and all the follow-up CBCTs.

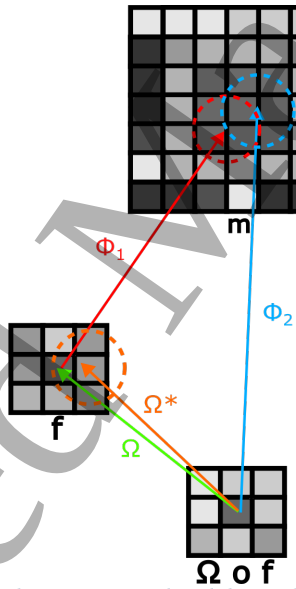



*Figure 2: Illustration of the method used to evaluate a normalized dense displacement field (DDF) distribution. $\boldsymbol{m}$ and $\boldsymbol{f}$ are the moved and fixed image respectively, $\Omega$ is a DDF generated by registering the fixed image to one of the follow-up image, $\Omega \circ \boldsymbol{f}$ is a virtual image generated warping $\boldsymbol{f}$ with a known DDF, $\boldsymbol{\phi_1}$ and $\boldsymbol{\phi_2}$ are the DDFs distribution created with the DIR method and $\Omega^*$ is the estimated DDF distribution which register $\boldsymbol{f}$ to $\Omega \circ \boldsymbol{f}$.*

## 2.6 Hyperparameter tuning

The hyperparameters of the supervised ($\alpha, \gamma, \omega$) and unsupervised ($\gamma, \omega$) models were fine-tuned in the validation dataset to achieve the best deformation capabilities while still predicting good-quality uncertainties based on the KL-Diverge metric. The hyperparameter tuning was performed iteratively. First, one hyperparameter was selected, and a grid search of the parameter that generated the best trade-off between the averaged DSC, MI, ECE, and the KL-Divergence metric on the validation dataset was selected. Then, this hyperparameter was fixed with its optimal value and the same procedure was repeated for the rest of the hyperparameters. To highlight the benefit in the image similarity, the MI was normalized (%MI) to be 0 when there is no registration and 100 when the MI is the lowest achieved by the model.

In this work, the value of $\sigma$ used to calculate $\boldsymbol{\Sigma}(\boldsymbol{\psi}, \sigma)$ was not considered a hyperparameter of the problem. Instead, the same value was used for all models. The smallest $\sigma$ which generated, on average, less than 0.1% of voxels with negative Jacobian in the sampled DDF ($\phi$) for all the models was selected to avoid over-smoothing the matrix $\boldsymbol{\Sigma}(\boldsymbol{\psi}, \sigma)$. We also fixed the average misalignment distance (d) used to calculate $\boldsymbol{\Sigma_p}$ to be equal to 2.4 mm. This value is within the expected misalignment error that can be found for different organs when using rigid registration in head and neck treatments [29]. $T(\boldsymbol{m})$ was calculated for each voxel as the maximum distance in all directions (x, y, z) that resulted in an average Hounsfield Unit (HU) difference of at least 20 HU compared to the voxel's position. We considered this HU threshold value since this corresponds to two sigmas of the average HU uncertainty of a CT.

## 2.7 Evaluation

The model's capabilities to predict the DDF and its uncertainty were evaluated in different ways. A qualitative assessment was performed by visually inspecting the DDF and its uncertainty. A semi-quantitative assessment was performed by looking at the trend of the DDF uncertainty when the spatial resolution of the pCT and the CBCTs gradually decreased. A quantitative evaluation was performed by comparing the evaluation metrics between the developed models and other DIR methods.

### Qualitative assessment

The supervised model capabilities to generate high-quality deformation and uncertainty estimation were visually evaluated by the authors to detect obvious gross miss-registrations in different challenging scenarios of the test set. The deformation capabilities were assessed in two patients who experienced significant anatomical changes throughout their treatment. The diagonal of the covariance matrix ($Diag(\boldsymbol{\Sigma}(\boldsymbol{\psi}, \sigma))$) was examined in regions where the wCT showed discrepancies with the CBCT, including cases where anatomical structures were present in one image but not in the other, as well as in the presence of image artifacts. Furthermore, the structure uncertainty was assessed by comparing probability maps of the binary target segmentations generated by a physician, B-spline and the dropout model.

### Aleatoric uncertainty analysis

The relation between the contrast of the input images and the DDF uncertainty, represented by $Diag(\boldsymbol{\Sigma}(\boldsymbol{\psi}, \sigma))$, was examined in the test set. Gaussian kernels with various standard deviations ($\hat{\sigma}$) were employed to smooth the input pairs (pCT, CBCT). The average of the median of the DDF uncertainty was calculated for all test samples.

### Model comparison

The performance of the models developed in this work were compared against three different models that can predict DIR uncertainty, a Monte Carlo dropout model (MCD), VoxelMorph model and a Monte Carlo B-spline model (MCBS) (See detailed description in the supplementary material).

The MCD model was implemented by introducing dropout layers in all the convolutional layers of the encoding block of the network. The weighted sum of the MI, BE, and Dice Similarity Coefficient Loss (DSCL) were used as the loss function [17], [19], [23]. The VoxelMorph model was implemented to predict a multivariate normal distribution of a dense displacement field [19]. The network of the model was a 3D-Unet which was extracted from the Pytorch version of their GitHub code. The loss function was the same as they proposed in their paper with the additional calculations developed in this paper for the term $\boldsymbol{LC}(\sigma)\boldsymbol{\Sigma}(\boldsymbol{\psi})\boldsymbol{C}(\sigma)$ of equation (1). The MCBS models was implemented using an open-source DIR algorithm (Elastix) that use MI and BE to regularize the displacement field. The Monte Carlo samplings were performed over the strength of the regularizer and the size of the B-spline grid. All these models hyperparameters were fine tuned in the

validation dataset to achieve similar mutual information than the supervised and unsupervised model. Then, all the models were compared in the test set considering the metrics described in section 2.5. For each metric, paired signed ranked Wilcoxon test was used to evaluate the median of the distribution of every model against the distribution of the supervised model. Bonferroni correction was used when multiple comparisons were performed.

# 3 Results

## 3.1 Hyperparameter tunning

An analysis over the smoothness of the DDF samples predicted with the models for different values of the standard deviation of the smooth convolution matrix $\boldsymbol{C}$ showed that using $\sigma_0 = 6$ voxels and a kernel size = 25 voxels generated on average less 0.1% of voxels with negative Jacobian. With these parameters, we used equation 4 to calculate the effective smooth convolution kernel $k(\sigma_0)$. Then we compared the effective kernel $k(\sigma)$ with the gaussian kernel (Figure 3).

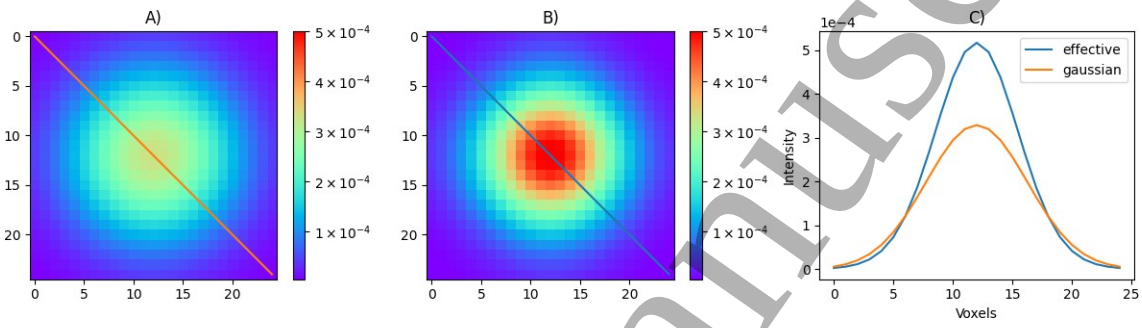



*Figure 3: Comparison of the middle slice of the normalized Gaussian kernel (A) and the normalized effective kernel (B) used in this work. C shows the diagonal profile indicated in the lines of images A and B.*

The effective kernel $k(\sigma)$ was sharper than the Gaussian kernel, giving more importance to neighboring voxels. With these parameters fixed we optimized the supervised and unsupervised network.

### Unsupervised Network

For the unsupervised model, it was found that $\frac{\gamma}{\omega}$= 50 generated the highest DSC in the limit of $\omega, \gamma \rightarrow \infty$. This also generated the lowest MI = -0.73, ie. %MI = 100, for all the explored combination of hyperparameters for this model. The maximum MI = -0.52, ie. %MI = 0, was achieved when no

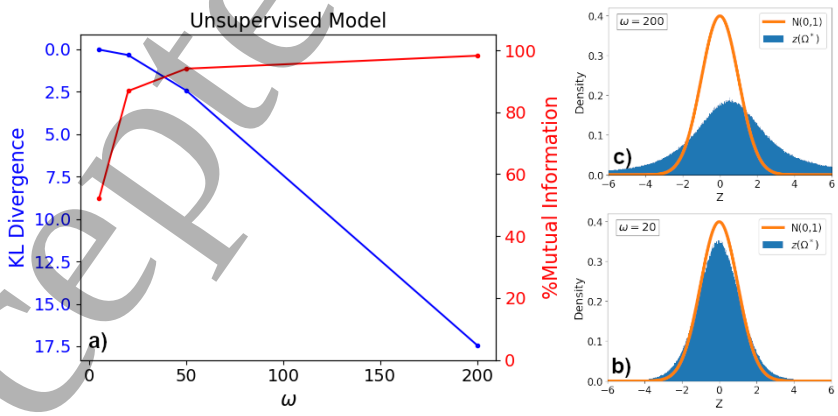



*Figure 4: a) shows the hyperparameter optimization plot for the unsupervised model with $\frac{\gamma}{\omega}$ = 50. The blue plot shows the average KL Divergence metric evaluated in the validation dataset as described in methods (lower is better). The axis is inverted to show lower KL divergence on the top. The red plot shows the average normalized mutual information evaluated in the validation dataset (higher is better image similarity). B and C compare the distribution $z(\Omega^*)$ (blue) with the $N(0,1)$ (orange) for a specific sample of the validation dataset and $\omega$ equal to 20 and 200 respectively.*

registration was applied. The optimization of $\omega$ showed that the DSC remained roughly stable at 0.855. However, the average normalized MI and the average KL-Diverge metric show a trade-off for different values of $\omega$ (Figure 4).

When $\omega$ = 20, the MI only decreased 15% from its maximum while the quality of the deformation field distribution has almost reached its maximum (KL-Divergence = 0). Moreover, the ECE analysis revealed a decline in conjunction with the KL Divergence metric for lower values of $\omega$ reaching its minimum (0.04) at ω = 20, as illustrated in Figure 5. As a compromise between registration accuracy and uncertainty reliability, $\omega$ = 20 was used for the supervised and unsupervised models.

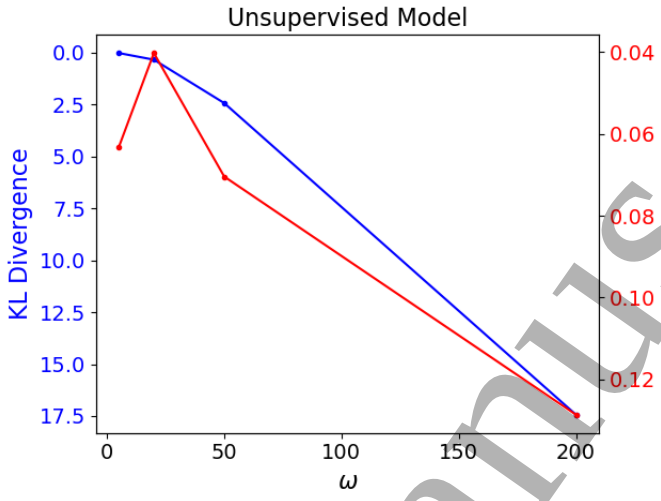



*Figure 5: Relation between the KL-Divergence metric developed in this work and the Expected Calibration Error (ECE) for different values of the hyperparameter ω of the unsupervised model.*

### Supervised Network

To fine-tune the hyperparameter $\alpha$ of the supervised network, the optimal hyperparameters of the unsupervised network ($\frac{\gamma}{\omega}$= 50 and $\omega$ = 20) were used. It was found that the MI and the KL vergence metric remained stable for different values of $\alpha$ but the DSC increased from 0.855 at $\alpha$ = 0 to 0.863 at $\alpha$ = 100. For this reason, in our following experiments, we used $\frac{\gamma}{\omega}$ = 50, $\omega$ = 20 and $\alpha$ = 100 for the supervised model.

## 3.2 Qualitative Assessment

The visual inspection of Figure 6 shows that the model could effectively predict a DDF that generate a wCT with similar characteristics to the CBCT even in presence of strong anatomical changes.

Because of the superior quality of the pCT, there are some anatomical structures that can be found on it but not in the daily CBCTs. In Figure 6, case 1, a blood vessel and muscle structures can be spotted in the pCT but they are not clearly seen in the CBCT. Because of this information loss, the model predicts high values in the $Diag(\boldsymbol{\Sigma}(\boldsymbol{\psi}, \sigma_0))$ in those regions.

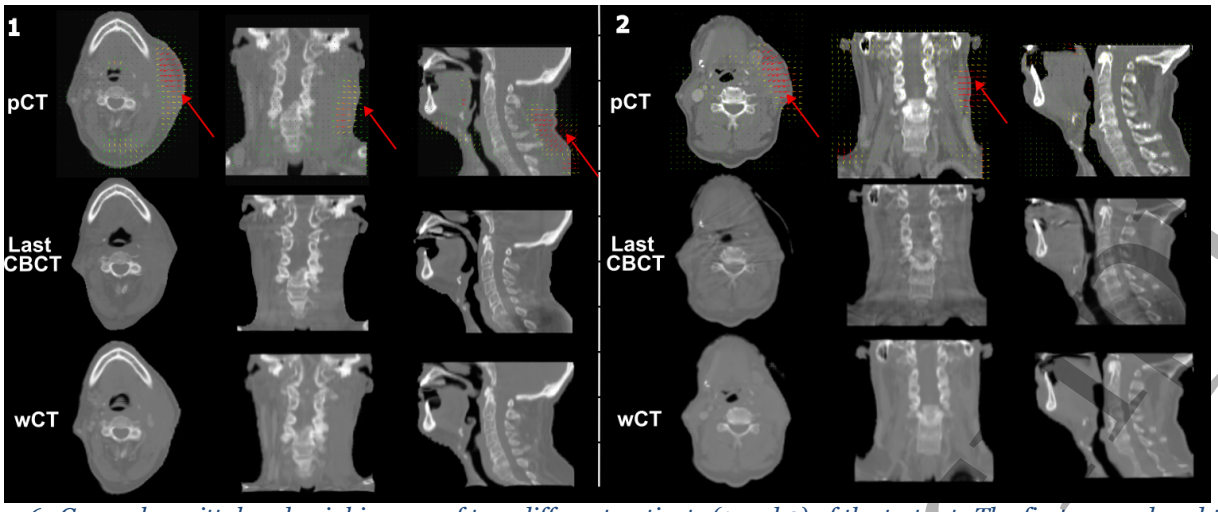



*Figure 6: Coronal, sagittal and axial images of two different patients (1 and 2) of the test set. The first, second and third rows show slices from the planning CT (pCT), the last CBCT of the treatment and the warped CT (wCT), propagated with the mean DDF, respectively. The DDF is illustrated with small arrows superimposed over different planes of the pCT. The color of the arrows qualitatively indicates the intensity of the DDF in the region, red for large, yellow and green for moderate and black for mild deformations. The large red arrows indicate the regions of the pCT which shrank during treatment.*

High uncertainty is also found in regions where the model could not find a proper DDF to register the pCT to the CBCT. In Figure 7, case 2, the anatomical differences in the oral cavity between the wCT and the CBCT are associated with an increase of the $Diag(\boldsymbol{\Sigma}(\boldsymbol{\psi}, \sigma_0))$. Image artifacts were also responsible for high DDF uncertainty. In Figure 7, case 3, the CBCT is cropped in the bottom right region of the pCT. Because of this loss of information, the DDF generates high uncertainty in that region. Note that high-contrast regions, such as bones, were generally associated with low uncertainty, and low-contrast regions in the pCT and CBCT were associated with high uncertainty. The magnitude of the DDF uncertainty also affects the confidence in the position of the warped structures. Figure 8 shows that the supervised model is confident ($PM \approx 1$) in regions where most binary structures agree and not confident in regions where the binary structures disagree ($PM \approx 0.5$).

## 3.3 Aleatoric uncertainty analysis

Figure 10 shows that the median of the DDF uncertainty of the supervised model significantly increases with the intensity of the smooth filter applied to the input images. This agrees with the observation in Figure 7, which shows that low-contrast regions have large DDF uncertainty. However, the VoxelMorph model doesn't associate significant changes in predicted uncertainty with the degree of smoothness and the dropout model predicts a smaller uncertainty when the contrast of the image decreases.

## 3.4 Model Comparison

The hyperparameters tunning of the dropout and VoxelMorph model also showed a similar trade-off as shown in Figure 4 (See results on supplementary material). The hyperparameters of these models were selected in such a way that these models generate the same mutual information for the validation dataset as obtained by the supervised and unsupervised models.

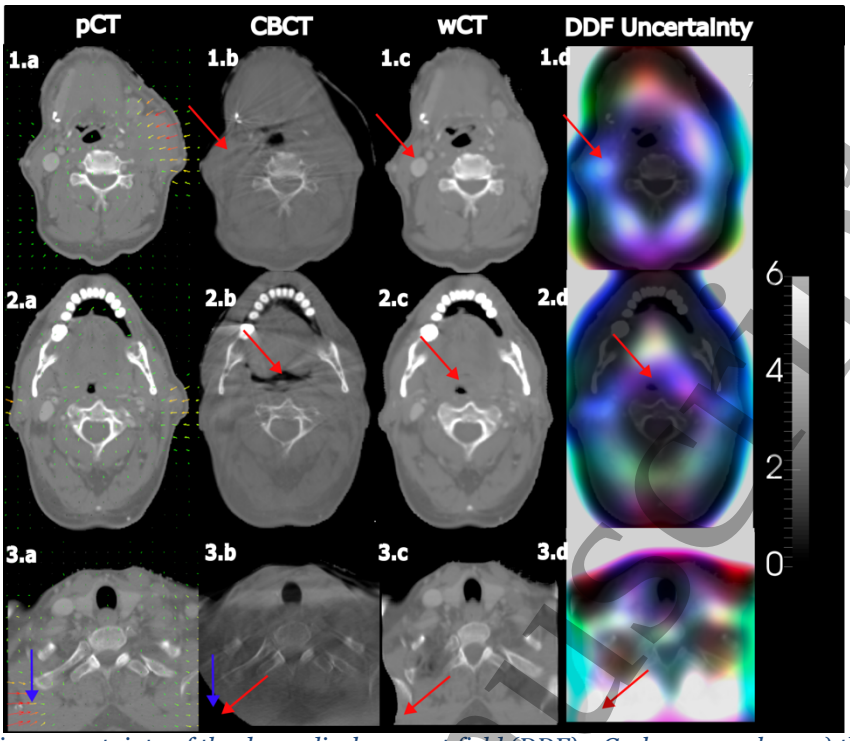



*Figure 7: Different images driving uncertainty of the dense displacement field (DDF). Columns show a) the planning CT (pCT), b) the CBCT, c) the warped CT (wCT), propagated with the mean DDF, and d) the diagonal elements of $\Sigma(\psi, \sigma_0)$ for different test samples in rows 1, 2 and 3. The DDF is illustrated with small arrows superimposed on the pCT. The color of the arrows indicates the intensity of the DDF in the region: red large yellow/green moderated and black mild deformation. The large red arrows show regions of the pCT that do not match the CBCT. The blue arrow shows a region of the patient's anatomy which is not capture in the CBCT. The vertical gray bar on the right is a scale for the intensity variations within the RGB image of $Diag(\Sigma(\psi, \sigma_0))$. Within this RGB representation, the uncertainty along the right-left direction is depicted in red, the uncertainty along the posterior-anterior direction is represented in green, and the uncertainty along the inferior-posterior direction is visualized in blue.*

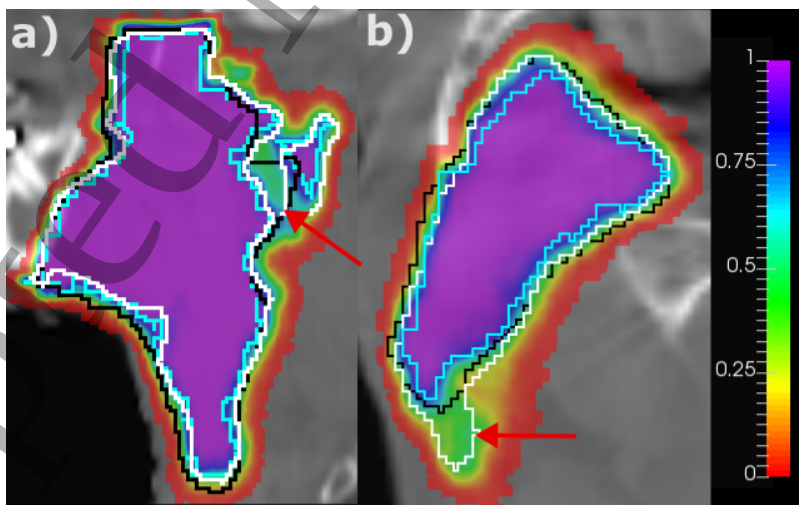



*Figure 8: Warped planning target volume for two different patients of the test set. The green contours were segmented by a physician; the white contours were propagated using the B-spline model, the cyan contours were propagated using the mean of the dense displacement field obtained using the dropout model and the probability maps (color map) were generated using sixty samples from the supervised model on a test set patient. The red arrow shows that regions in which the binary contours do not match each other present low confidence in the probability map ($PM \approx 0.5$).*

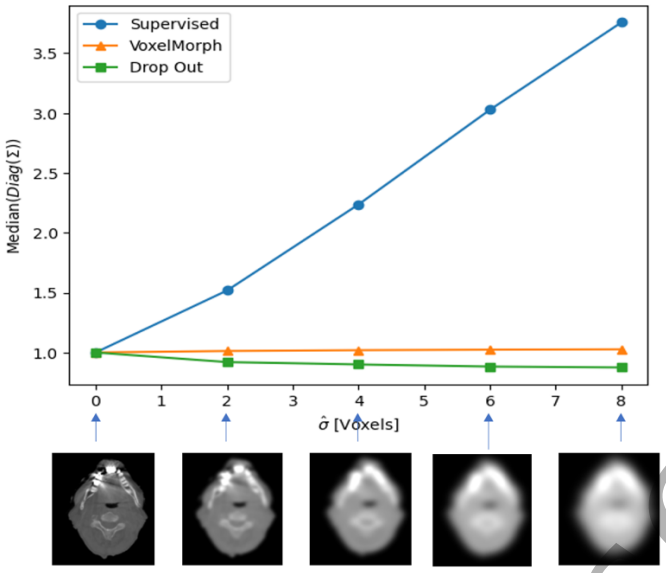



*Figure 10: Dependence of the dense displacement field uncertainty with the image contrast. The y-axis shows the normalized average median Diag($\Sigma(\psi, \sigma_0)$) calculated for all the test samples. Because of difference in the magnitude of the uncertainty predicted with the different methods, the median uncertainty values were normalized with their value at $\hat{\sigma}$=0. The x-axis the standard deviation of the Gaussian filter applied to the input image pairs. The small images at the button of the figure show an example of the CBCTs smoothed with $\hat{\sigma}$.*

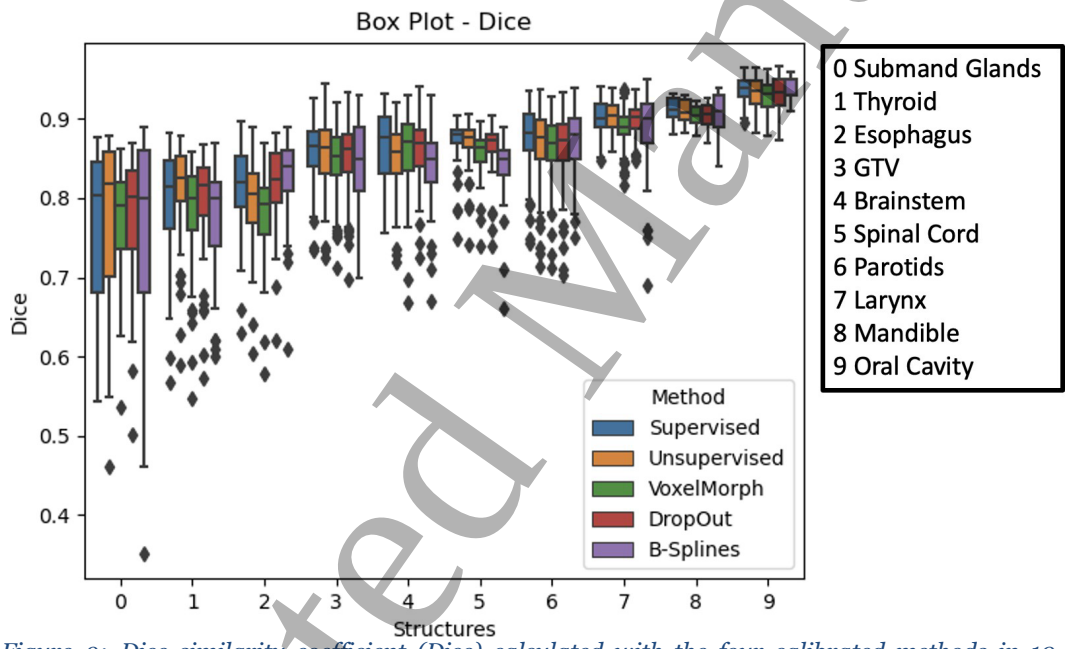



*Figure 9: Dice similarity coefficient (Dice) calculated with the four calibrated methods in 10 different organs of the test dataset.*

Even though the dice similarity coefficient for all the test contours and all the models don't present obvious visible differences (Figure 9), the supervised model exhibited statistically superior performance ($p < 0.005$) compared to the unsupervised model in 6/10 assessed organs (Organs: 2, 4, 5, 6, 8, 9). At the same time, it demonstrated statistically inferior performance in the submandibular glands and the thyroid when compared to the unsupervised model. In contrast, when compared to the dropout model, the unsupervised model displayed statistically better performance in 5/10 (3, 5, 6, 8, 9), while no statistically significant differences were observed in the remaining organs. A similar pattern was observed when comparing the supervised model with the B-spline model, as the supervised model exhibited superior performance in organs 1, 3, 4, 5, and 6, while no statistical differences were observed in the other organs. Notably, the unsupervised model outperformed the VoxelMorph model in all the assessed organs.

| Model | DSC ↑ | MI ↓ | KL Divergence ↓ | ECE ↓ | run Time (s) ↓ |
|---|---|---|---|---|---|
| Supervised | 0.891 [0.850, 0.920] | -62.98 [-73.10, -59.55] | 00.11 [00.08, 00.14] | 0.03 [0.02, 0.04] | ~ 2 |
| Unsupervised | 0.888 [0.848, 0.917] | -63.03 [-73.61, -59.79] | 00.15 [00.11, 00.21] | 0.03 [0.02, 0.04] | ~ 2 |
| VoxelMorph | 0.880 [0.833, 0.912] | -65.03 [-73.51, -60.70] | 21.35 [16.44, 26.92] | 0.17 [0.14, 0.19] | ~ 1.5 |
| Dropout | 0.883 [0.846, 0.915] | -64.55 [-74.00, -60.48] | 21.95 [16.08, 26.89] | 0.16 [0.14, 0.18] | ~ 120 |
| B-spline | 0.892 [0.844, 0.925] | -60.19 [-69.85, -55.65] | - | 0.18 [0.14, 0.21] | ~ 7661 |

*Table 1: Comparison of different deformable image registration models that can predict DIR uncertainty. The first value shows the median of the metric over all scenarios and the brackets show the first and third quartile of parameter distribution. DSC is the dice similarity coefficient between the structures propagated with μ and the physician's structures (higher better). The MI is the average mutual information between 60 different samples of the warped CTs and the CBCT (lower better). The KL divergence corresponds to the metric developed in this work to evaluate the quality of the DIR distribution (lower is better). The ECE is the average expected calibration error calculated between the probability maps for all the structures (lower better). The run time is the time spent by the computer to obtain the uncertainty of the DDF.*

Although the supervised model demonstrated enhanced performance in specific organs, with a higher DSC compared to both the unsupervised model and the B-spline model, there were no statistically significant differences ($p < 0.05$) in terms of the average DSC across all organs (Table 1). This suggests that while the supervised approach may excel in specific organ segmentation tasks, its overall performance, as measured by the average DSC across all organs, remains comparable to that of the B-spline and unsupervised models. However, the supervised model outperformed both the dropout and VoxelMorph models as measured by the DSC.

When analyzing the mutual information, no statistically significant differences were observed between the supervised model and the dropout, unsupervised, and VoxelMorph models. This can be attributed to the hyperparameter tuning, which aimed to achieve comparable image similarity performance to assess the uncertainty estimation capabilities fairly.

When evaluating the ECE of probability maps and KL divergence metric, the two models presented in this study demonstrated superior performance compared to the other models. Additionally, in terms of computational efficiency, our models outperformed all the models except for VoxelMorph. This discrepancy arises from the nature of the algorithms employed. Our model and VoxelMorph directly estimate uncertainty from the network, whereas the other models need sampling over various parameters, resulting in time-consuming computations.

# 4 Discussion

In this work, we built a probabilistic multi-resolution image registration model that uses prior and image information to estimate the parameters of a normal distributed dense displacement field (DDF) using convolutional neural networks (CNNs). The model offers the flexibility to be trained in either an unsupervised or supervised manner, enabling the integration of segmentation probability maps to enhance the accuracy of DDF distribution prediction. Our model has shown comparable or superior performance when compared to other DIR models considering metrics related to DDF accuracy, such as mutual information (MI) and dice similarity coefficient (DSC), and has outperformed all the methods when considering metrics that evaluate the DDF distribution quality, such as the KL-Divergence and expected calibration error (ECE). The analysis carried out in this paper has shown that our model can generate registration in a matter of seconds, allowing fast and reliable deformable image registration.

The assumption that the propagated structures follow an independent Bernoulli distribution generates a cross-entropy (CE) term in the loss function of the supervised model. This metric has been extensively used in segmentation models but has not been used to optimize a DIR distribution before [30], [31] . The integration of CE into our model offers distinct advantages over the use of the DSC employed in previous works [12], [19], particularly in terms of providing a more probabilistic interpretation of the contour matching metric and improving the calibration of the probability maps [24]. In summary, the supervised model presented in the paper has a significant advantage

over other methods, as it not only allows to use segmentations provided by a physician but also to use segmentations coded as probability maps, where probability maps reflect aggregated overlap of contours drawn by multiple evaluators.

Unlike other evaluation methods that use DDFs derived from plausible anatomical deformations to assess the DDF accuracy [27], [28], our approach provides a way to perform a probabilistic evaluation of the whole DDF distribution in a multimodal image registration problem. Notably, we observed a correlation between this metric and the ECE (Figure 5), which is a metric that indirectly assesses the quality of the DDF by looking at the calibration of the probability maps. However, our metric presents some limitations since it only works for multivariate normal DDF distributions and requires the generation of likely anatomical deformations.

The hyperparameter tuning of the deep learning models employed in this study revealed a trade-off between the average M, calculated with different realizations of the wCT and the CBCTs, and the quality of the DDF distribution. This trade-off arises due to the fine-tuning of the hyperparameter that governs the amplitude of the DDF dispersion. Higher DDF dispersions can impact voxel-wise matching in regions with high-intensity gradients, resulting in reduced image similarity. These trade-off curves are characteristic of the model performance to predict DIR uncertainty since better models will thoroughly increase the uncertainty in regions of the image where this will not significantly affect the image similarity (low contrast areas). Our unsupervised and supervised model achieved the best trade-off balance between these two metrics, allowing the generation of a DDF distribution with KL-Divergence of 0.2 with an average normalized Mutual Information (nMI) of 85% during validation. Even though we selected $\omega = 20$ to generate these values, the optimal value of this hyperparameter should be selected based on how much accuracy the user is willing to sacrifice to gain confidence in the uncertainty estimated. This trade-off is particular to the problem at hand.

The visual inspection of the DDF uncertainty showed that our model presents high uncertainty in smoother regions of the images (aleatoric uncertainty) and in regions where the matching of different anatomical structures was not performed accurately (epistemic uncertainty). The uncertainty estimated with our model might also present characteristics for out-of-distribution detection (OOD). In the experiment conducted in Figure 10, the use of over-smoothed images can be regarded as an exploration of OOD scenarios. By observing the model's response to such a simplistic analysis, we could suspect that it could recognize inputs that fall outside the distribution it was trained on. The heightened uncertainty for these extreme cases indicates that the model exhibits greater uncertainty when encountering unfamiliar or abnormal inputs, such as over-smoothed images. However, a more detailed analysis should be performed to make a stronger claim.

Despite the theoretical advantages of the supervised model over the unsupervised model, the analysis of the DSC revealed superior performance in only 6 out of 10 organs. Moreover, the supervised model did not exhibit better results when assessing the average DSC across all organs. We suspect that the absence of ground truth segmentation during the training of the supervised model limited its ability to generate even better contour propagation. This problem is significantly more relevant in small organs such as the submandibular glands and the thyroid, in which subtle differences between the predicted and validation mask generate bigger changes in the DSC. This might explain why the unsupervised model performed better in these structures. On the other hand, it is expected that overall the Dropout model and the supervised model did not present statistical differences in DSC because both methods are supervised. Oppositely, the supervised model achieved better DSC than VoxelMorph because the former was trained in a supervised way.

The overall mutual information (MI) did not show statistical differences for the learning-based models because we set the loss function weights to generate similar MI between these models to make a fair analysis of the DDF distribution. The quality of the DDF distribution was relevant only for the unsupervised and supervised models developed in this work. The main reason why the VoxelMorph model could not predict relevant DDF distribution while generating a high image

similarity is because, in the loss function, the model couples the uncertainty term with the regularization term using the hyperparameter $\lambda$. The dropout model's lack of performance in predicting a good DIR distribution could be interpreted as a limitation of the model in estimating the aleatoric uncertainty (Figure 10). As Kendall et al. [32] described, such models cannot estimate the uncertainty related to noise in the input data. Although we could not assess the quality of the DIR distribution of the Monte Carlo B-spline model, its performance can be estimated to be poor because of its high probability map ECE. However, we believe these results might be influenced by a poor selection of the phase space in investigating B-spline uncertainty.

Incorporating structures within the supervised model inevitably introduces a bias in estimating DDF uncertainty when compared to the unsupervised model. While quantifying the exact magnitude of this bias presents challenges, the assessment of uncertainty prediction consistency, as gauged by KL divergence, reveals statistically significant differences (Wilcoxon $p < 0.05$). Specifically, the supervised model demonstrates an average KL divergence of 0.11, whereas the unsupervised model registers a slightly higher average of 0.15. This statistically significant difference in uncertainty prediction between the two methods suggests a bias introduced by the supervised model. Incorporating multiple physician segmentations during the training process could increase the bias in the supervised model, thereby generating better DIR distribution prediction in the vicinity of the structures.

One of the most direct applications of this method in clinics is in the area of adaptive radiotherapy since the estimation of DIR uncertainty can be used to assess better the dose accumulated over the treatment fractions quantitatively and to decide if re-planning is needed [3]. Furthermore, fast and reliable DIR can be used in daily online adaptive radiotherapy for contour propagation. The advantage of predicting the structure uncertainty might open the possibility to speed up the adaptive workflow by incorporating these uncertainties in the treatment optimizer and thus reduce the need for the time-consuming structure physician reviewing.

# 5 Conclusion

In conclusion, we developed a probabilistic multi-resolution image registration model that uses convolutional neural networks to estimate the dense displacement field distribution between two multimodal images. Compared to other models that can estimate DIR uncertainty, our model presented equal or better contour propagation capabilities, predicted a higher quality DDF distribution and presented faster execution times. Future research will study the consequence of the DIR uncertainty when re-planning in daily adaptive radiotherapy workflow.

# 6 Acknowledgments

This project has received funding from the European Union's Horizon 2020 Marie Skodowska-Curie Actions under Grant Agreement No. 955956.

# References

[1] E. Vargas-Bedoya, J. C. Rivera, M. E. Puerta, A. Angulo, N. Wahl, and G. Cabal, "Contour Propagation for Radiotherapy Treatment Planning Using Nonrigid Registration and Parameter Optimization: Case Studies in Liver and Breast Cancer," *Applied Sciences*, vol. 12, no. 17, 2022, doi: 10.3390/app12178523.

[2] G. E. Christensen, J. H. Song, W. Lu, I. El Naqa, and D. A. Low, "Tracking lung tissue motion and expansion/compression with inverse consistent image registration and spirometry," *Med Phys*, vol. 34, no. 6Part1, pp. 2155–2163, 2007, doi: https://doi.org/10.1118/1.2731029.

[3] I. J. Chetty and M. Rosu-Bubulac, "Deformable Registration for Dose Accumulation," *Semin Radiat Oncol*, vol. 29, no. 3, pp. 198–208, 2019, doi:

https://doi.org/10.1016/j.semradonc.2019.02.002.
[4] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast MR images," *IEEE Trans Med Imaging*, vol. 18, no. 8, pp. 712–721, 1999, doi: 10.1109/42.796284.
[5] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, "elastix: A Toolbox for Intensity-Based Medical Image Registration," *IEEE Trans Med Imaging*, vol. 29, no. 1, pp. 196–205, 2010, doi: 10.1109/TMI.2009.2035616.
[6] J. Shackleford *et al.*, "Plastimatch 1.6 – current capabilities and future directions," Oct. 2012, pp. 108–119.
[7] B. Hunt *et al.*, "Fast Deformable Image Registration for Real-Time Target Tracking During Radiation Therapy Using Cine MRI and Deep Learning," *International Journal of Radiation Oncology*Biology*Physics*, vol. 115, no. 4, pp. 983–993, 2023, doi: https://doi.org/10.1016/j.ijrobp.2022.09.086.
[8] F. Albertini, M. Matter, L. Nenoff, Y. Zhang, and A. Lomax, "Online daily adaptive proton therapy," 2020.
[9] J. Luo *et al.*, "On the Applicability of Registration Uncertainty," 2019, pp. 410–419. doi: 10.1007/978-3-030-32245-8_46.
[10] M. Hub, M. L. Kessler, and C. P. Karger, "A Stochastic Approach to Estimate the UncertaintyInvolved in B-Spline Image Registration," *IEEE Trans Med Imaging*, vol. 28, pp. 1708–1716, Oct. 2009, doi: 10.1109/TMI.2009.2021063.
[11] A. J. Smolders, A. J. Lomax, D. C. Weber, and F. Albertini, "Deep learning based uncertainty prediction of deformable image registration for contour propagation and dose accumulation in online adaptive radiotherapy," *Phys Med Biol*, 2023, [Online]. Available: http://iopscience.iop.org/article/10.1088/1361-6560/ad0282
[12] Y. Hu *et al.*, "Weakly-supervised convolutional neural networks for multimodal image registration," *Med Image Anal*, vol. 49, pp. 1–13, 2018, doi: https://doi.org/10.1016/j.media.2018.07.002.
[13] G. Balakrishnan, A. Zhao, M. R. Sabuncu, A. V Dalca, and J. Guttag, "An Unsupervised Learning Model for Deformable Medical Image Registration," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9252–9260. doi: 10.1109/CVPR.2018.00964.
[14] T. Che *et al.*, "AMNet: Adaptive multi-level network for deformable registration of 3D brain MR images," *Med Image Anal*, vol. 85, p. 102740, 2023, doi: https://doi.org/10.1016/j.media.2023.102740.
[15] M. Magris and A. Iosifidis, "Bayesian learning for neural networks: an algorithmic survey," *Artif Intell Rev*, vol. 56, no. 10, pp. 11773–11823, 2023, doi: 10.1007/s10462-023-10443-1.
[16] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," *Proceedings of The 33rd International Conference on Machine Learning*, Oct. 2015.
[17] X. Gong, L. Khaidem, W. Zhu, B. Zhang, and D. Doermann, "Uncertainty Learning towards Unsupervised Deformable Medical Image Registration," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 1555–1564. doi: 10.1109/WACV51458.2022.00162.
[18] N. Cheng, O. Malik, S. Becker, and A. Doostan, "Bi-fidelity Variational Auto-encoder for Uncertainty Quantification." Oct. 2023.
[19] A. V Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces," *Med Image Anal*, vol. 57, pp. 226–236, 2019, doi: https://doi.org/10.1016/j.media.2019.07.006.
[20] W. Weng and X. Zhu, "UNet: Convolutional Networks for Biomedical Image Segmentation," *IEEE Access*, vol. 9, pp. 16591–16603, 2021, doi: 10.1109/ACCESS.2021.3053408.
[21] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE Trans Med Imaging*, vol. 22, no. 8, pp. 986–1004, 2003, doi: 10.1109/TMI.2003.815867.
[22] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "Deep learning in medical image registration: a review," *Phys Med Biol*, vol. 65, no. 20, p. 20TR01, Oct. 2020, doi: 10.1088/1361-6560/ab843e.
[23] O. Weistrand and S. Svensson, "The ANACONDA algorithm for deformable image registration in radiotherapy," *Med Phys*, vol. 42, no. 1, pp. 40–53, 2015, doi: https://doi.org/10.1118/1.4894702.
[24] A. Mehrtash, W. Wells, C. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence Calibration

and Predictive Uncertainty Estimation for Deep Medical Image Segmentation." Oct. 2019.
[25] J. Pukala, S. L. Meeks, R. J. Staton, F. J. Bova, R. R. Mañon, and K. M. Langen, "A virtual phantom library for the quantification of deformable image registration uncertainties in patients with cancers of the head and neck," *Med Phys*, vol. 40, no. 11, 2013, doi: 10.1118/1.4823467.
[26] K. Nie, C. Chuang, N. Kirby, S. Braunstein, and J. Pouliot, "Site-specific deformable imaging registration algorithm selection using patient-based simulated deformations," *Med Phys*, vol. 40, no. 4, 2013, doi: 10.1118/1.4793723.
[27] S. R. van Kranen, S. van Beek, C. Rasch, M. Herk, and J.-J. Sonke, "Setup Uncertainties of Anatomical Sub-Regions in Head-and-Neck Cancer Patients After Offline CBCT Guidance," *Int J Radiat Oncol Biol Phys*, vol. 73, pp. 1566–1573, Oct. 2009, doi: 10.1016/j.ijrobp.2008.11.035.
[28] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "A review of deep learning based methods for medical image multi-organ segmentation," *Physica Medica*, vol. 85, pp. 107–122, 2021, doi: https://doi.org/10.1016/j.ejmp.2021.05.003.
[29] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation BT - Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015," N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241.
[30] R. Boyd, A. Basavatia, and W. A. Tomé, "Validation of accuracy deformable image registration contour propagation using a benchmark virtual HN phantom dataset," *J Appl Clin Med Phys*, vol. 22, no. 5, pp. 58–68, May 2021, doi: 10.1002/acm2.13246.
[31] G. Loi *et al.*, "Performance of commercially available deformable image registration platforms for contour propagation using patient-based computational phantoms: A multi-institutional study: A," *Med Phys*, vol. 45, no. 2, pp. 748–757, Feb. 2018, doi: 10.1002/mp.12737.
[32] A. Kendall and Y. Gal, "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 5580–5590.