DEGREE PROJECT IN MEDICAL ENGINEERING,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2020

Deep Learning Based Deformable Image Registration of Pelvic Images

BLANCA CABRERA GIL

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ENGINEERING SCIENCES IN CHEMISTRY, BIOTECHNOLOGY AND HEALTH

Deep Learning Based Deformable Image Registration of Pelvic Images

BLANCA CABRERA GIL

Master in Medical Engineering
Date: June 3, 2020
Supervisor: Jonas Söderberg
Examiner: Matilda Larsson
School of Engineering Sciences in Chemistry, Biotechnology and Health
Host company: RaySearch Laboratories AB
Swedish title: Bildregistrering av bäckenbilder baserade på djupinlärning

Abstract

Deformable image registration is usually performed manually by clinicians, which is time-consuming and costly, or using optimization-based algorithms, which are not always optimal for registering images of different modalities. In this work, a deep learning-based method for MR-CT deformable image registration is presented. In the first place, a neural network is optimized to register CT pelvic image pairs. Later, the model is trained on MR-CT image pairs to register CT images to match its MR counterpart.

To solve the unavailability of ground truth data problem, two approaches were used. For the CT-CT case, perfectly aligned image pairs were the starting point of our model, and random deformations were generated to create a ground truth deformation field. For the multi-modal case, synthetic CT images were generated from T2-weighted MR using a CycleGAN model, plus synthetic deformations were applied to the MR images to generate ground truth deformation fields. The synthetic deformations were created by combining a coarse and fine deformation grid, obtaining a field with deformations of different scales.

Several models were trained on images of different resolutions. Their performance was benchmarked with an analytic algorithm used in an actual registration workflow. The CT-CT models were tested using image pairs created by applying synthetic deformation fields. The MR-CT models were tested using two types of test images. The first one contained synthetic CT images and MR ones deformed by synthetically generated deformation fields. The second test set contained real MR-CT image pairs. The test performance was measured using the Dice coefficient. The CT-CT models obtained Dice scores higher than 0.82 even for the models trained on lower resolution images. Despite the fact that all MR-CT models experienced a drop in their performance, the biggest decrease came from the analytic method used as a reference, both for synthetic and real test data. This means that the deep learning models outperformed the state-of-the-art analytic benchmark method. Even though the obtained Dice scores would need further improvement to be used in a clinical setting, the results show great potential for using deep learning-based methods for multi- and mono-modal deformable image registration.

Sammanfattning

Bildregistrering görs vanligtvis för hand eller med optimeringsbaserade algoritmer, vilket är tidskrävande och kostsamt. I detta arbete presenteras en djupinlärningsbaserad metod för icke-linjär registrering av MR bilder mot CT bilder. Först optimeras ett neuralt nätverk för att registrera par av CT-bilder av bäcken. Senare tränas modellen på MR-CT-bildpar för att registrera CT-bilder mot dess MR-motsvarighet.

Lämplig ground-truth data för detta problem saknas vilket löses med två tillvägagångssätt. I fallet med par av CT-bilder var utgångspunkten identiska bilder där en av dessa sedan deformeras med ett slumpmässigt genererat deformationsfält innan bilderna matades till nätverket. I det multimodala fallet genererades syntetiska CT-bilder från T2-viktad MR med användning av en CycleGAN-modell. Dessutom applicerades syntetiska deformationer på MR-bilderna för att generera deformationsfält för ground-truth. De syntetiska deformationerna skapades genom att kombinera ett grovt och fint deformationsnät, vilket gav ett fält med deformationer i olika skalor.

Flera modeller tränades på bilder med olika upplösningar. Deras resultat jämfördes med en analytisk algoritm som används i ett faktiskt arbetsflöde för bildregistrering. CT-CT-modellerna testades på bildpar skapade med syntetiska deformationsfält. MR-CT-modellerna testades på två typer av testbilder. Den första innehöll syntetiska CT-bilder och MR-bilder deformerade av syntetiska deformationsfält. Den andra testuppsättningen innehöll riktiga MR-CT-bildpar. Testprestanda mättes med hjälp av Dice-koefficienten. Resultaten visade att CT-CT modellerna erhöll Dice-koefficient högre än 0,82 även för modellerna tränade på bilder med lägre upplösning. Trots det faktum att prestanda minskade för alla MR-CT-modeller, kom den största minskningen från den analytiska metoden som användes som referens, både för syntetisk och verklig testdata. Detta innebär att djupinlärningsmodellerna överträffade den analytiska benchmarkmetoden. Även om de erhållna Dice-koefficienterna skulle behöva förbättras innan användning i en klinisk miljö, visar resultaten att djupinlärningsbaserade metoder för multi- och monomodal bildregistrering har stor potential.

Acknowledgements

In the first place, I would like to thank Jonas Söderberg for his help and guidance throughout this project, as well as to Stina Svensson and Ola Weistrand for sharing their expertise and knowledge on image registration. I would also like to thank Iridium Kankernetwerk for providing the anonymized patient data that has been utilized in this project. Additionally, I would like to express my gratefulness to Chunliang Wang for his feedback and improvement ideas. Finally, I would like to thank my family for their unconditional support.

List of Abbreviations

CT: Computed Tomography

MR: Magnetic Resonance

DIR: Deformable Image Registration

ROI: Region of Interest

Contents

Chapter 1

Introduction

According to the American Cancer Society (ACS), the most predominant types of cancer among American men aged over 55 years old are prostate and bladder cancer. The 5-year survival rate for patients diagnosed with prostate cancer is 100% if the disease is only in the prostate and nearby organs. However, this figure drops to 30% if the cancer has spread to other parts of the body. Similarly, the 5-year survival rate for bladder cancer is 77%. If the tumor is invasive but has not yet spread outside the bladder the 5-year survival rate is 69%, but if the cancer has extended to the surrounding tissue or to nearby lymph organs this survival rate drops to 35% [5] [3]. These figures highlight the importance of obtaining an early diagnose of the disease and perform an accurate treatment plan.

The main imaging modality for radiation therapy planning and dose computation is computed tomography (CT) scan. The poor contrast that characterizes CT images makes it very challenging to obtain an accurate segmentation of target structures and tumors. On the other hand, magnetic resonance (MR) images show excellent soft-tissue contrast but do not provide the electron density information needed for dose computation. Therefore, MR images are used together with CT images to achieve target and tumor delineation. An accurate delineation of these images is crucial for a correct radiotherapy plan and dose delivery [29].

Image registration is used in the medical field to match images acquired from different viewpoints, at different times, containing physiological variations and/or obtained using different scanning modalities [16]. Combining multiple images in this way can be used to quantify changes in organ shape, size, and

position, providing physicians a better understanding of the patient's anatomy and organ function [19]. Moreover, the establishment of the correspondence between images is critical to a wide variety of clinical tasks such as image fusion, organ atlas creation, and tumor growth monitoring [16]. Additionally, the application of deformable registration in image-guided radiotherapy provides improved geometric and dosimetric accuracy of radiation treatments [19].

Traditionally, cross-modality image registration is performed manually by clinicians. As a consequence, the final registration is highly dependent on the expertise of the user and very costly. Automatic methods based on analytic algorithms have also been developed. A commonly used cost function is mutual information (MI) which measures the reduction in uncertainty of one image given the knowledge of another [9]. The main problem faced when registering CT to MR images is that the later ones do not possess a calibrated intensity scale. This means that images obtained from different scanners usually have different intensity scales and probability distributions, resulting in MI getting stuck in local maxima when the images' intensity scales are very different [29].

The arrival of deep learning methods has allowed to obtain state-of-the-art results in many computer vision tasks including image registration. However, most results for deformable image registration using deep learning are recent and in practice the problem is still solved by analytic methods.

1.1 Aims

Deformable image registration plays a key role for accurate treatment planning. It is used by clinicians to propagate contours and map dose definitions between image sets. This task is important for an efficient workflow and to avoid manually contouring of regions of interest. As stated in Section 1, automated algorithms using analytical methods are not always optimal when performing deformable registration between multi-modal images [8] [16]. Therefore there is a need to find a better solution.

The main aim of the study is to assess the viability of a deep learning model to perform the deformable registration task. Later, the obtained results will be compared with an analytic method that is being used in an actual registration workflow.

One of the main challenges that will be faced along this project is the lack of

available image registration ground truth data. Correct ground truth registrations are usually not available since they have to be created by hand, which is a time consuming and expensive process. This problem has been addressed in the literature in two different ways. The first one is using a similarity metric as loss function during training [14]. However, this approach is not completely adequate for multi-modal registration as the similarity metric can converge to a local maximum. The second approach is to generate synthetic ground truth deformation fields [12]. In this project, the second approach is being implemented.

The main goal of the project is to develop a deep learning-based model for multi-modal deformable image registration for the male pelvic region. In the literature, deep learning has been mostly used to solve the deformable registration problem for images of the same modality. Thus, this work is considered as a study on the viability of using a neural network for multi-modal image registration. In order to achieve this goal two subgoals are set:

- Develop a model to register CT images to synthetically deformed CT images. This model is going to be evaluated on synthetic test deformation fields.
- Train a model to register synthetic CT images to match MR images based on the results obtained in the previous step. This model is going to be evaluated both on synthetic test deformations and real images.

Chapter 2

Methods

The method for image registration being investigated in this project follows the work of [12]. A convolutional neural network is used to predict a deformation field given a reference and a target image. The network is a modified 3D U-net [27] with two input channels, one for each 3D image, and three output channels for the x-, y- and z-components of the deformation field. Synthetic deformations are generated to create reference and target image pairs for training since, as previously mentioned, ground truth deformation fields are not available. More precisely, random synthetic deformation fields are applied to training images, yielding pairs of reference and target images. The image pairs are then fed through the network which results in predicted deformation fields. The predictions are compared with the synthetic deformations and a loss is computed. A graphic representation of the method is presented in Figures 2.1 and 2.2. The network is trained by minimizing a loss function. The method has a number of interesting hyperparameters:

- Resolution of input images.
- Method for generating synthetic deformation fields.
- Architecture of the convolutional neural network.
- Resolution of the predicted deformation field.

The resolution of input images and the network architecture will be discussed in Section 3.1 and 3.2, and the method for generating synthetic deformation fields in Section 2.4. The method presented here uses the same resolution for the predicted fields as for the input images since this simplifies the network architecture. The same choice is made in [12].

Compared to the work of [12], which is concerned with the registration of pulmonary CT images, this work faces the additional difficulty presented by cross-modality registration. Applying the method of [12] to register MR and CT images requires perfectly aligned image pairs as training data. Such aligned data is normally not available and therefore synthetically generated CT images will be used for training the network. The generation of synthetic CT images is described in Section 2.4.

Firstly, a CT to CT registration model is created and evaluated to make sure that the results from [12] can be transferred to CT images of the pelvic region. Secondly, the MR-CT registration model is investigated.

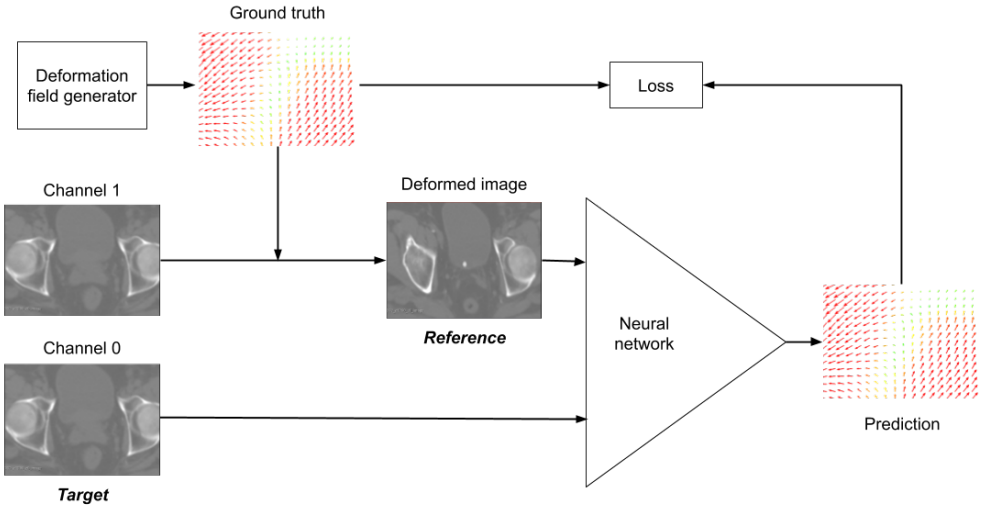


Figure 2.1: Representation of the implemented method for CT-CT registration. Two identical images are the starting point of the pipeline. A ground truth synthetic deformation vector field is generated and applied to the image in channel 1 to generate a reference image from which the deformation field to obtain is known. Then, both images, reference and target are fed to the network and a deformation field is predicted. Finally, the loss is calculated by comparing the ground truth and the predicted deformation fields.

Throughout the project, the terms reference and target images are going to be used. The term *reference image* refers to the stationary image, while by *target image* it is meant the image to be transformed to be mapped to the reference image.

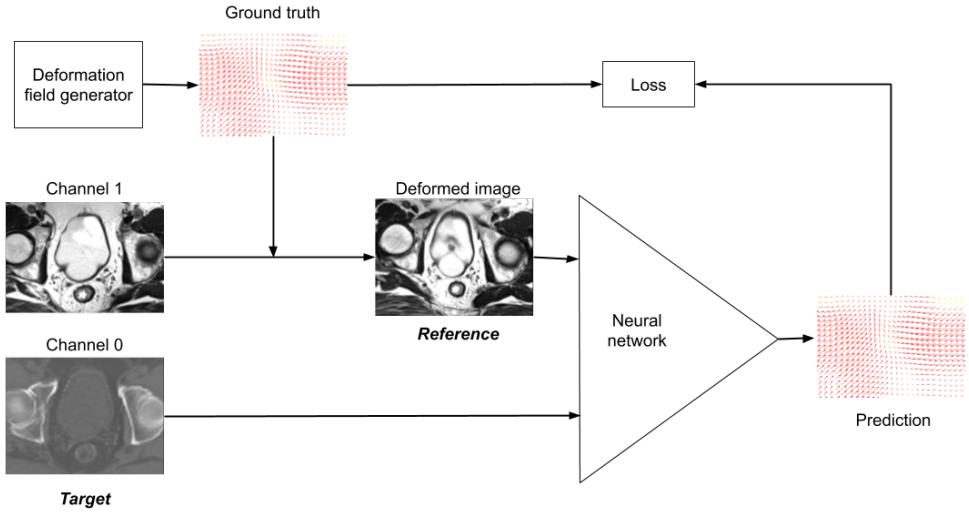


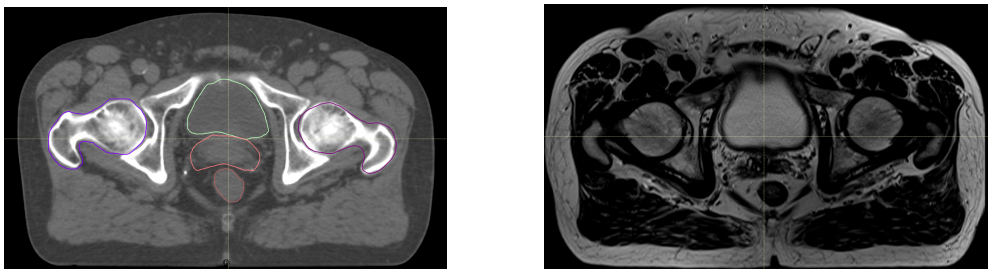
Figure 2.2: Representation of the implemented method for MR-CT registration. Two perfectly aligned images are the starting point of the pipeline: a T2 MR image and a synthetic CT. A ground truth synthetic deformation vector field is generated and applied to the image in channel 1 to generate a reference image and a ground truth deformation field. Then, both images, reference and target are fed to the network and a deformation field is predicted. Finally, the loss is calculated by comparing the ground truth and the predicted deformation fields.

2.1 Dataset

To train, test and validate our model two different datasets were used, both containing MR and CT images of the male pelvic region. The first one is from Iridium Kankernetwerk, Antwerp, Belgium, and was used for training, validating and testing the models. The second one, the Gold Atlas research dataset from [25] was used as test set for the MR-CT models.

2.1.1 Iridium

The iridium dataset has a total of 425 anonymized patients containing different MR sequences, CT and Cone Beam CT examinations of the male pelvic region. The CT exams contain clinically approved and peer-reviewed contours, that were used in delivered radiotherapy plans. In order to get the images that fit the purpose of the project, only patients with CT and T2 MR examinations were selected. T2-weighted MR images are used for radiation therapy because



(a) CT contoured image. (b) T2 MR image.

Figure 2.3: Sample patient data from Iridium database.

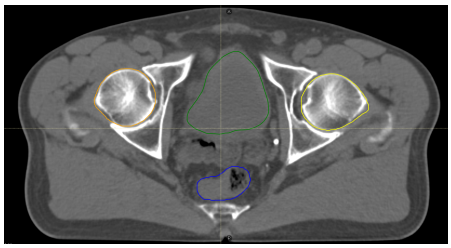
they brighten tissues containing fat and water which allows to detect pathologies [15]. On top of that, the CT examinations were required to have contours for the bladder, right and left femur, prostate, and rectum. These regions of interest (ROIs) will be used to monitor training and evaluate the model. Additionally, patients that had a hip prosthesis were removed from the dataset. After this selection, a total of 186 patients were left. To ensure a homogeneous distribution of the data in the training, validation, and test sets, the samples were randomly split into 38 test samples, 20 validation samples, and 128 training samples. An example of the image data can be seen in Figure 2.3.

2.1.2 Gold Atlas

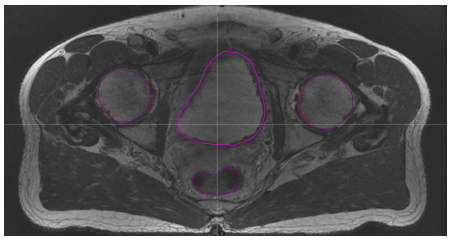
The Gold Atlas dataset is presented in [25] as a way to provide a dataset for the training and validation of segmentation algorithms. The dataset contains T1- and T2-weighted MR images as well as CT images of 19 patients in the same positions with multi-observer and expert consensus delineations of relevant organs of the male pelvic region. The contours relevant for our purposes are the bladder, rectum, prostate, and femur bones. Since the CT images did not have any delineations in this dataset, such were created with an existing automatic segmentation tool: the deep learning model Iridium Pelvic Male of RayStation system was used. This dataset was only used to test the accuracy of the multi-modal deep learning registration models. An example of the available data in the Gold Atlas dataset can be found in Figure 2.4.

2.2 Data Preprocessing

The original images from both datasets had varying image shapes and voxel sizes. For this reason, all images were resampled to a predetermined resolution



(a) CT with deep learning based segmentations.



(b) T2 MR image with consensus segmentations.

Figure 2.4: Sample patient data from Gold Atlas database.

and size before being presented to the neural network. This was done both during training and at inference. In most experiments, the input images were cropped to a physical size of (23.0, 15.0, 20.0) cm. This size was chosen to ensure that most of the bladder, rectum, prostate, and most of the femoral heads would fit in the image for a typical patient. The reason for keeping a smaller image size was that there was a limit on the total number of voxels that the neural network could operate on, so a smaller image size enabled the use of a higher resolution. The limit on the number of voxels is due to the fact that the neural network had to fit into GPU-memory. It is worth mentioning that parts of the femoral heads often ended up outside of the images as seen in Figure 2.5. The voxel sizes for the different models trained in this work are presented in Table 2.1. The model Iridium MR-CT complete (see Table 2.1) is using an image resolution of (0.25, 0.25, 0.25) cm to be able to use the images' complete field of view.

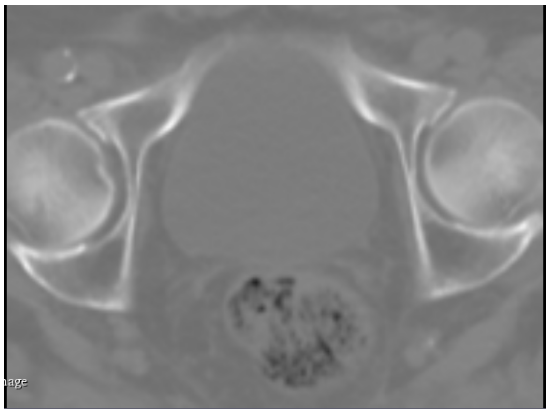


Figure 2.5: Example of cropped CT Image of the Iridium database.

Test	Voxel Size (cm)	Image Shape (voxels)
Iridium CT-CT	(0.3, 0.107, 0.107)	(76, 140, 186)
Iridium CT-CT	(0.3, 0.144, 0.144)	(160, 104, 138)
Iridium CT-CT/ MR-CT	(0.25, 0.25, 0.25)	(92, 60, 80)
Iridium CT-CT	(0.5, 0.5, 0.5)	(46, 30, 40)
Iridium MR-CT	(0.4, 0.084, 0.084)	(58, 176, 236)
Iridium MR-CT all image	(0.25, 0.25, 0.25)	(82, 80, 146)
Gold Atlas	(0.25, 0.097, 0.097)	(92, 154, 204)

Table 2.1: Data resolution and corresponding image shapes.

2.3 Data Augmentation

Data augmentation was used to increase the amount of training data. Every time that a data sample was fed to the network, it was transformed by applying a set of deformations on the fly. In this way, the network never saw the same input twice. Data augmentation is a type of regularization and prevents overfitting. The applied augmentations were combinations of rotations, translations, and elastic deformations. Translation and rotation values were picked from a uniform random distribution with boundaries +/- a given value. The elastic deformations were created by picking random displacement vectors from a normal distribution on a coarse grid and creating intermediary displacement vectors by spline interpolation. The distributions used for creating data augmentations are shown in Table 2.2.

Parameter	Value
Translation (cm)	0.5
Rotation (deg)	2
Grid spacing (cm)	(10, 10, 10)
Deformation scale (cm)	(0.1, 0.1, 0.1)

Table 2.2: Values for the random translation, rotation and deformation scale used for data augmentation.

2.4 Synthetic ground truth and CT generation

Due to the lack of available ground truth data for the registration task, synthetic data was used. In the following section, the methods and choice of parameters for generating synthetic data are presented.

2.4.1 Ground Truth Generation

As mentioned earlier, the models were trained using perfectly aligned image pairs, where the reference image was created by applying a known deformation field to one of the images in the pair. Such reference-target image pairs can be easily generated if a large pool of deformation fields are available. One of the assumptions in this project is that suitable deformation fields can be generated by a fairly simple process and that there is no need for them to be anatomically correct.

The pelvic region is characterized by having organs that can experience completely different types of deformations. On one hand, the bones only suffer from rigid-body transformations, while the rectum and the bladder can experience a great increase in size in a very short time. Accordingly, it was decided to concatenate a coarse and a fine deformation grid to train the network with fields of different characteristics. This was also the approach used in [12]. The coarse grid allows the network to learn how to register big deformations, while the fine grid allows the network to learn smaller ones. To generate a deformation field, a grid spacing parameter was chosen at random from a given interval of values. The set of deformation vectors of each deformation field was obtained from a random uniform distribution having boundaries at +/- a determined deformation scale parameter. The choice of parameters used to generate the ground truth fields can be found in Table 2.3. Also, Figure 2.6 is a representation of deformation parameter's meaning. These parameters were chosen after testing different kinds of deformations and selecting the ones which resembled examples of real deformations. To avoid translations, the resulting field was normalized. Once the coarse and fine fields were obtained, they were concatenated and the resulting deformation field was saved as ground truth, applied to the channel 1 input and its corresponding label map. An example of a generated deformation field can be found in Figure 2.7.

The choice of parameters of the synthetic deformation fields are very impor-

tant given that they have a great influence on the network's learning and its capability to perform well when seeing real data.

Parameter	Fine Grid	Coarse Grid
Grid spacing (cm)	[2,3]	[7,15]
Deformation scale (cm)	0.2	1.5

Table 2.3: Choice of parameters to generate the ground truth vector fields.

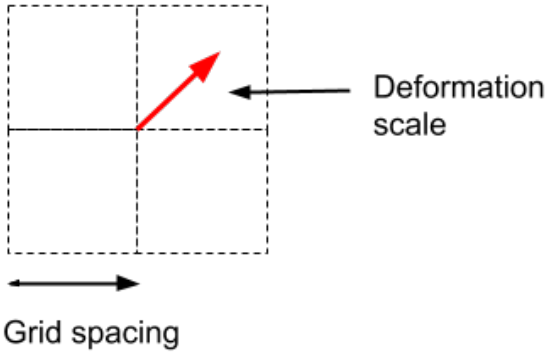


Figure 2.6: Graphic representation of the deformation parameters and its meaning.

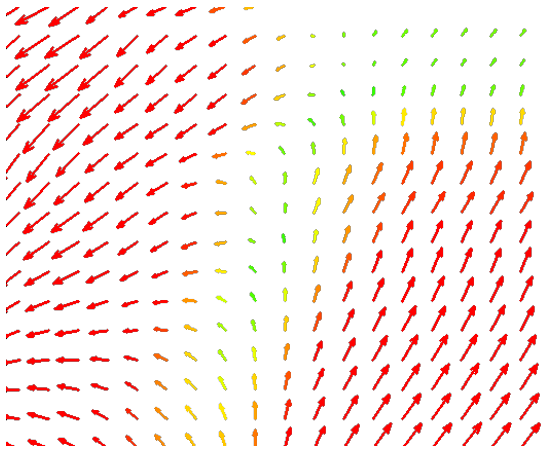
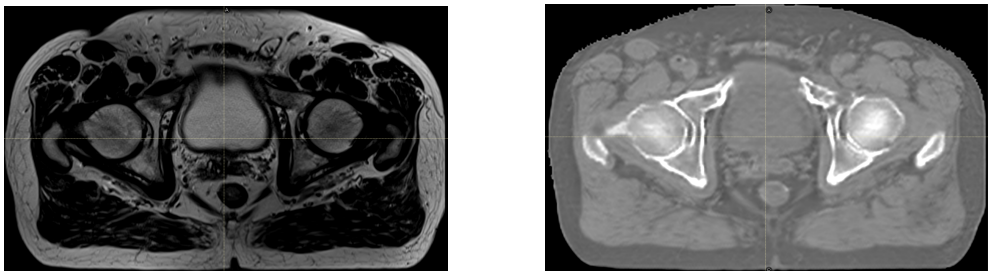


Figure 2.7: Example of generated ground truth deformation field.



(a) T2 MR image. (b) Synthetic CT image.

Figure 2.8: Example of resulting synthetic CT image from a T2 MR image using CycleGAN.

2.4.2 Synthetic CT generation

In order to tackle the multi-modal deformable image registration (DIR) problem it is necessary to have MR-CT image pairs with a corresponding ground truth deformable registration. In the case of this study, this data is not available, therefore synthetic CT images were generated to solve the problem. Using synthetic images for DIR in this way has been previously done by [29] and [11]. The procedure implies to use a CycleGAN architecture [18] previously trained to generate synthetic CT (sCT) images from T2-weighted MR. In our case, there was no need to train a CycleGAN network as it was already done as part of another project. The benefit of generating the input images in such a way is that the resulting image pair is already perfectly aligned and ready to be fed into the network. On the other hand, it also implies the risk that the network may not generalize when facing real data. The T2 MR images to be transformed are part of the Iridium dataset. An example of a resulting image is presented in Figure 2.8.

2.5 Neural Network architecture

The network architecture proposed in this project is based on the one presented in [12]. It is a modified version of U-net used to solve the problem of mono-modal deformable registration field estimation. In their work, four main modifications were introduced to the original network. The first one was to feed the network with two inputs: the target and the reference images. Secondly, the architecture was deepened one more level. Also, the activation functions were changed from ReLU to Leaky ReLU. Finally, the output convolutional layer of the network was changed to have three feature maps, one for each di-

mension (x,y,z) of the vector field to be predicted. The graphic representation of the network architecture proposed in [12] can be found in Figure A.5. In [12], the neural network was optimized to solve the registration problem for lung images. Thus, a set of hyperparameter optimization has been conducted to improve the performance of the network when facing images of the pelvic region. More details about the hyperparameter optimization can be found in Sections 2.6 and 3.1.

2.6 Hyperparameter optimization

In order to find the best hyperparameter configuration that allowed the network to obtain the greatest performance on validation data, a grid search was conducted. The learning rate, number of epochs, optimizer, loss function, number of convolutions per block, number of layers and their number of filters, the usage of residual connections, and the input image resolution were the parameters to be tuned during the optimization. The different hyperparameter configurations tested during the search can be found in Table 2.4. For all the tests an image patch size of (23.0, 15.0, 20.0) cm was used as described in Section 2.2. The different tests were ordered depending on their run time. In this way, tests with (0.5, 0.5, 0.5) image resolution were performed in the first place due to its lower computation time. After, the 3 best performing configurations were tested on (0.25, 0.25, 0.25) resolution images. From these results, the best performing configuration was selected and tested on (0.144, 0.144, 0.144) and (0.3,0.107, 0.107) resolution images. The approximate training time to complete 2000 epochs was different for each model. The 0.5 resolution models lasted about a day, the 0.25 resolution ones about 4 days, and the higher resolution ones about 3 weeks.

The addition of residual connections in the network architecture is presented in Figure 2.9. The residual connection adds the input of the convolutional block to the result of batch normalization before the last activation function of the convolutional block. This arrangement is depicted in Figure 2.10.

Parameter	Configurations
Learning Rate	1, 0.1, 0.5, 0.05, 0.001
Epochs	150, 700, 1500, 5000
Optimizer	Adagrad, Adam, Adadelta
Loss Function	MSE, L2, MAE
Convolutions per block	4, 2
Filters per layer	(32,64,128,256,512), (32,64,128,256,512,1024)
Residual	True, False
Resolution	(0.5,0.5,0.5), (0.25,0.25,0.25), (0.144,0.144,0.144), (0.3, 0.107,0.107)

Table 2.4: Hyperparameter optimization configurations, where MSE is mean squared error and MAE mean absolute error.

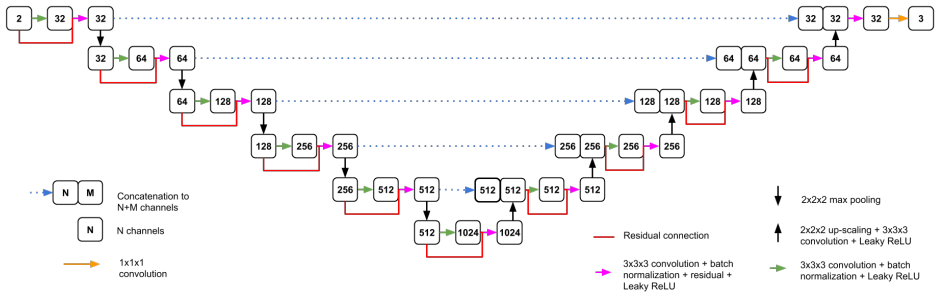


Figure 2.9: Network architecture implemented in this project. It is based in the network presented in [12], but it has been deepened one more layer and residual connections have been added in each convolutional block.

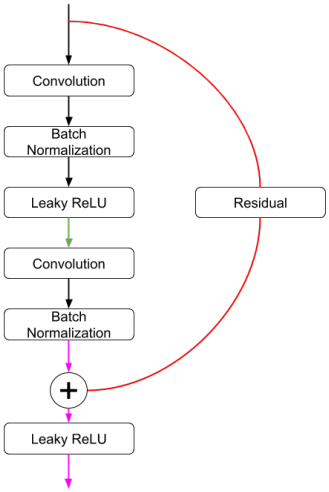


Figure 2.10: Residual block architecture used in the neural network. This diagram represents in greater detail the meaning of the green, red and purple arrows of Figure 2.9

2.7 Training and Evaluation metrics

In this section, the training and evaluation metrics used in the neural network are going to be presented as well as its formulas. The notation used in the equations is the following: a represents the ground truth deformation field, b the predicted vector field, i, j, k are the vector components for each dimension and n is the total number of training samples.

After performing the hyperparameter grid search and testing the model performance for different loss functions the one which provided better results was *L2* loss. Its formula is stated in Equation 2.1.

$$L2loss = \sum_{i=1}^{n} (a - b)^2 \tag{2.1}$$

To monitor the evolution of the accuracy of the predicted vector fields during training, three main measures were used. In the first place, the mean euclidean error between the ground truth and the predicted deformation field is monitored throughout the epochs. Its formula can be found in 2.2. The second metric that was monitored during training was the mean error relative to the average displacement of the deformation field. It is calculated as stated in

Equation 2.4 and 2.3.

$$EuclideanError = \sqrt{(a_i - b_i)^2 + (a_j - b_j)^2 + (a_k - b_k)^2} \tag{2.2}$$

$$Displacement = \sqrt{(a_i)^2 + (a_j)^2 + (a_k)^2} \tag{2.3}$$

$$RelativeError = \frac{EuclideanError}{Displacement} \tag{2.4}$$

Additionally, to evaluate the performance of the registration on specific regions of interest during training, the Dice coefficient was also monitored. The Dice coefficient is an overlap measure often used to quantify the similarity between two binary regions. The classical Dice coefficient is defined as in Equation 2.5 [26].

$$DC = \frac{2\,|A \cap B|}{|A| + |B|} \tag{2.5}$$

This coefficient was used to compare the ground truth labels with the deformed ones by applying the predicted deformation field. This way, the accuracy of the prediction can be sensed in a more reliable way. The Dice coefficient was monitored during training, validation, and testing for the following ROIs: right femoral head, left femoral head, bladder, rectum, and prostate.

2.8 Implementation

The model was implemented using Python as programming language and Tensorflow 1.12 as the machine learning library to build the neural network. CUDA 9.0 was used as the parallel computing platform. The trainings have been executed on a GPU-server with NVIDIA Tesla V100-SXM2-32GB GPUs.

Chapter 3

Experiments & Results

3.1 Hyperparameter optimization

The best performing hyperparameters are shown in Table 3.1. Also, in Table 3.2, a comparison of the Dice scores and standard deviations between the baseline and the optimized model are presented. The baseline model is an implementation of the neural network presented in [12].

Parameter	**Configurations**
Learning Rate	0.1
Epochs	4000
Optimizer	Adagrad
Loss Function	L2
Convolutions per block	2
Filters per layer	(32, 64, 128, 256, 512, 1024)
Residual	True

Table 3.1: Best performing hyperparameters.

Model	Metric	R Femur	Bladder	Rectum	L Femur	Prostate
Baseline	Dice	0.85	0.92	0.85	0.85	0.84
	Std	0.06	0.02	0.03	0.07	0.07
Optimized	Dice	0.88	0.95	0.89	0.83	0.89
	Std	0.05	0.01	0.03	0.09	0.04

Table 3.2: Average Dice score and standard deviation comparison between the baseline architecture from [12] and the best model from the hyperparameter search on validation data.

Additionally, different interpolation schemes for synthetic deformations were investigated. To apply a deformation, the image was resampled using an interpolation method. Spline interpolation of first order resulted in more blurred images compared to applying third-order splines. Accordingly, three different strategies were considered:

1. Always interpolate with splines of order 3.
2. Switch at random between first and third-order splines.
3. Deform 90% of training images applying first and third-order interpolation at random, and fed the remaining non-deformed 10% mixed in-between the deformed samples.

The three strategies were evaluated at the end of the hyperparameter search being the third the most successful one. The results of all the different configurations tested during the grid-search can be found in Table B.1.

3.2 Experiments

After obtaining the results of the hyperparameter search, there was one more parameter that needed to be explored. This was the training image resolution. Therefore, several models were trained on images of different resolutions for CT-CT and MR-CT registration to assess the one that provided a better performance. For all the experiments, the test data was also registered using ANACONDA algorithm to be able to benchmark the results from the deep learning-based models. ANACONDA algorithm is the analytic method that is used nowadays in RayStation software (RaySearch Laboratories AB, Stockholm, Sweden), it is described in more detail in Section A.3.

Firstly, the tests on CT-CT image registration are presented in Section 3.2.1. After, the test on MR-CT image registration are presented in Section 3.2.2.

3.2.1 CT - CT models

To asses the influence of the image resolution during training, four different models were trained. These had the same network configuration but were trained on images of different resolutions. In this case, the image resolution was also considered a hyperparameter. The resolutions in which the models were trained are:

- (0.5, 0.5, 0.5) cm
- (0.25, 0.25, 0.25) cm
- (0.144, 0.144, 0.144) cm
- (0.3, 0.107, 0.107) cm

The reason behind the choice of resolutions was to see how the performance of the model was affected when the information in the images was reduced. It was interesting to see how the anisotropy of the resolution would affect the learning of the network for each dimension. During these tests, the output deformation vector field was resampled to meet full image resolution (0.3, 0.107, 0.107) cm. In this way, the performance of the model was tested on the original image resolution. The metric used to compare the results between the different models was the Dice coefficient per organ. The results are presented in Figure 3.1, where the Dice distribution between the reference and target images is represented in blue, the Dice after registering the images with ANACONDA is represented in orange and labeled as RS, the registrations from the model trained on full image resolution is labeled as 0107, the results from the model trained on isotropic full image resolution is named 0144, and the registrations from the models trained on isotropic image resolutions 0.25 and 0.5 are labeled as 025 and 05 respectively. An example of the resulting deformations can be found in Figure 3.2. The learning curves for the 0107 model are presented in Figure 3.3.

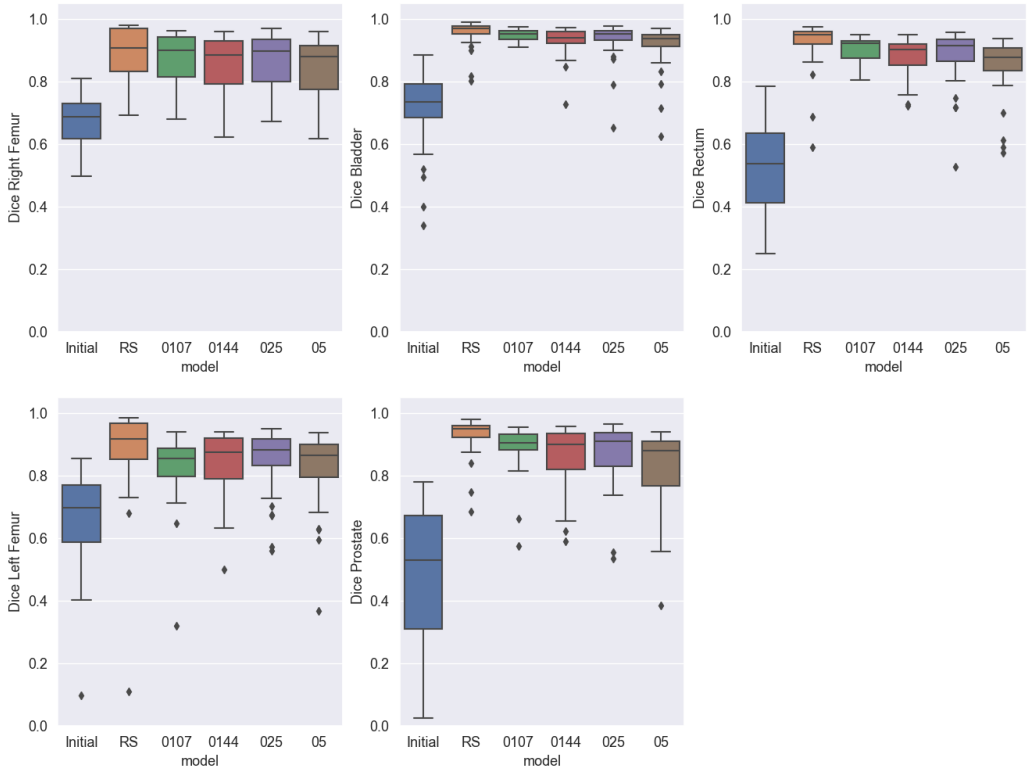
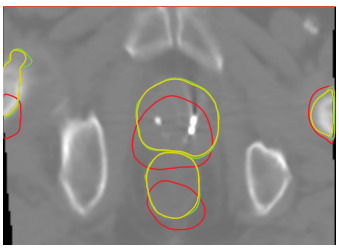
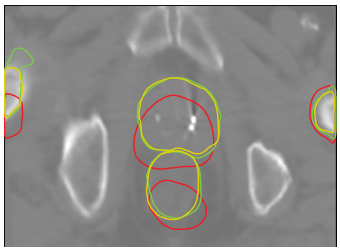


Figure 3.1: Resulting Dice coefficient score per organ of the different CT-CT trained models on test images. RS represents the results obtained with ANACONDA algorithm, 0107 is the model trained on full resolution images, 0144 is the model trained on isotropic full resolution images, 025 is the model trained on 0.25cm resolution images, and 05 is the model trained on 0.5cm resolution images.

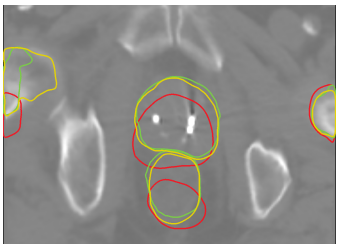
In Table 3.3 the mean Dice scores of every model per organ are presented as well as the results obtained from ANACONDA algorithm (RS) for the same dataset. For comparison, Dice scores from a deep learning segmentation model, also trained on the Iridium dataset, are included. The model was validated in [7] and its segmentations were found to be acceptable with no or minor corrections in the majority of the cases. These Dice scores will be referred to as benchmark scores from now on. When comparing these scores to the ones obtained from our model, it can be seen that the femoral heads' scores are slightly lower than the benchmark ones. As mentioned before, it can be explained by the fact that in some images these ROIs are cropped which increases the difficulty of its registration. Nevertheless, for the bladder, rectum, and prostate regions the obtained Dice scores are very similar or higher to the benchmark



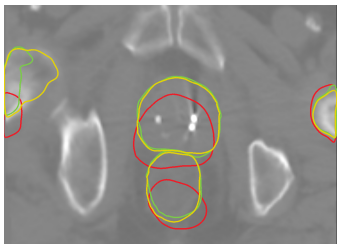
(a) Deformation result from ANACONDA



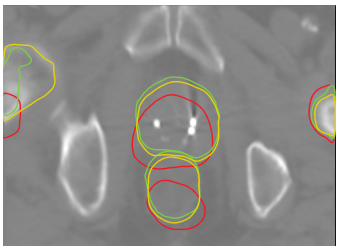
(b) Deformation result from 0107



(c) Deformation result from 0144



(d) Deformation result from 025



(e) Deformation result from 05

Figure 3.2: Comparison of the final registrations obtained by the different CT-CT models on the same test patient of the Iridium dataset. The background image is the deformed target image. In red the initial non-deformed masks, in green the reference masks, and in yellow the predicted deformed masks. The masked organs that appear in the images are both femoral heads, the rectum and the prostate.

ones even for the models trained on low-resolution images.

3.2.1.1 Error Analysis

To better understand how the differences in the initial deformations can influence the outcome of the network, the relationship between the ground truth displacement size and the prediction error is analyzed. Further analysis of the behavior of the models when facing different kinds of deformation fields can be found in Section B.2. In Figure 3.4 the relationship between the prediction

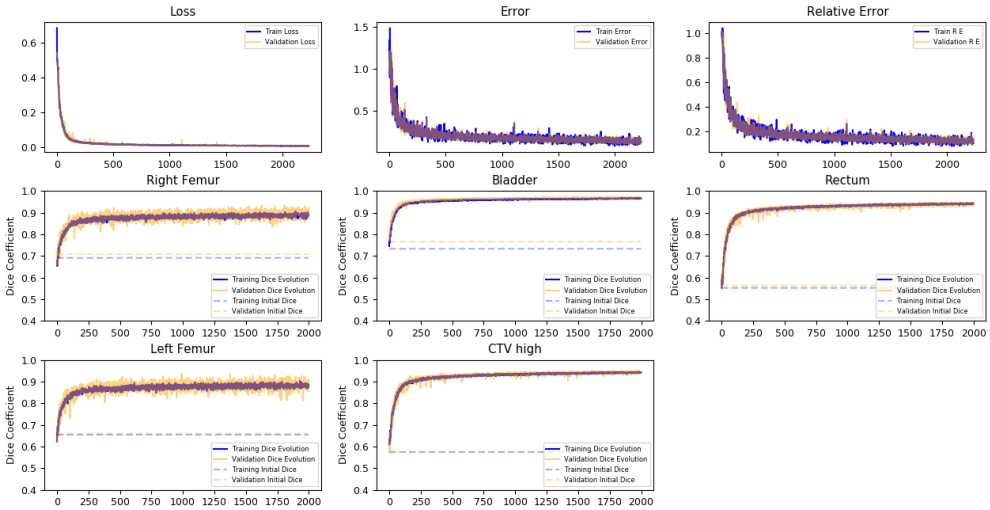


Figure 3.3: Training and validation curves of the CT-CT model trained on full resolution images. The training is monitored with the loss, euclidean error between the predicted and ground truth deformation fields, relative error to the average displacement, and Dice scores per organ. Blue represents the metrics for training data, while orange represents the metrics for validation data. The dashed lines depict the initial Dice scores.

Model	Metric	R Femur	Bladder	Rectum	L Femur	Prostate
0107	Dice	0.87	0.94	0.89	0.82	0.88
	Std	0.07	0.01	0.04	0.1	0.07
0144	Dice	0.85	0.93	0.88	0.84	0.86
	Std	0.08	0.04	0.05	0.09	0.09
025	Dice	0.86	0.93	0.88	0.84	0.87
	Std	0.08	0.05	0.08	0.09	0.09
05	Dice	0.84	0.91	0.85	0.82	0.82
	Std	0.09	0.06	0.08	0.1	0.12
RS	Dice	0.89	0.95	0.92	0.87	0.92
	Std	0.07	0.03	0.07	0.14	0.05
Benchmark scores	Dice	0.94	0.93	0.90	0.94	0.82

Table 3.3: Average Dice scores and standard deviations per model per organ for the CT-CT experiments. The benchmark scores are from a deep learning segmentation model trained on the same dataset [7]

error and the average displacement for 100 randomly selected image voxels is

presented. From this figure, it can be seen that the greater the displacement the more likely it is to obtain a greater prediction error, yet, there is not a straight linear relationship.

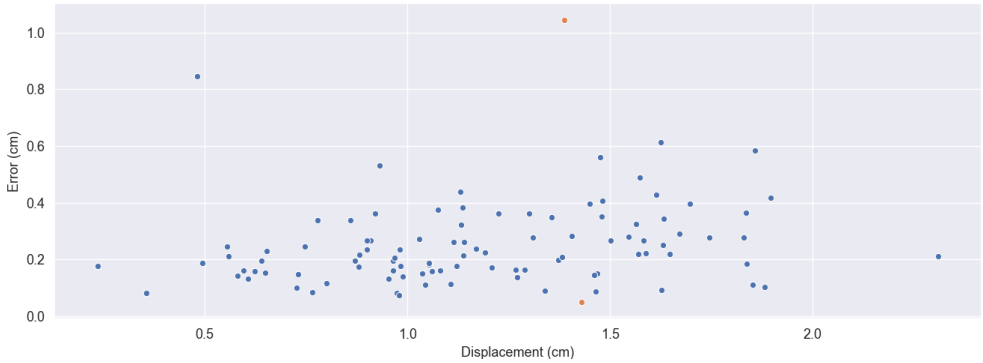


Figure 3.4: Error-Displacement analysis. In this figure, the relationship between the displacement of the ground truth field and the error of the prediction is presented. In the plot, a 100 randomly selected voxels of the model's 0107 test results are analyzed. This means that individual voxels are being evaluated. The minimum and maximum error measurements are colored in orange.

In the same way, the error map showing the average prediction error over all test samples is presented in Figure 3.5. The error presents 3 main interesting behaviors. In the first place, there are low error regions creating a wavy pattern. Secondly, a moderate source of error is located in the middle of the region where the bladder, prostate, and rectum are most likely to be located. Finally, the left and bottom image borders present a high source of errors.

3.2.2 MR - CT models

For the multi-modality model no hyperparameter optimization was performed due to time constraints. Nevertheless, different training image configurations were tested to see which one gave a better outcome. Therefore, 3 models were trained:

- Full MR resolution (0.4, 0.084, 0.084) with cropped images.
- Low isotropic resolution (0.25, 0.25, 0.25) with cropped images.
- Low isotropic resolution (0.25, 0.25, 0.25) with complete images.

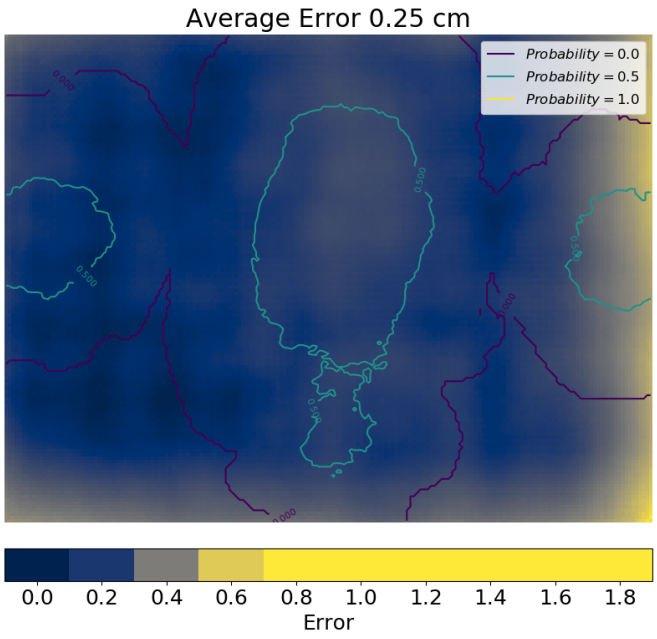


Figure 3.5: Average prediction error per voxel of the 0107 CT-CT model. The value of each image pixel is the average error over all the test set. The contours state the probability that a ROI is outside the stated region. On top of the image the overall average error is stated.

The later image configuration was proposed to see the effect of using all the image information in the learning process. However, as the available training space was limited, the image resolution was lowered. The training evolution of the full resolution model can be found in Figure 3.6.

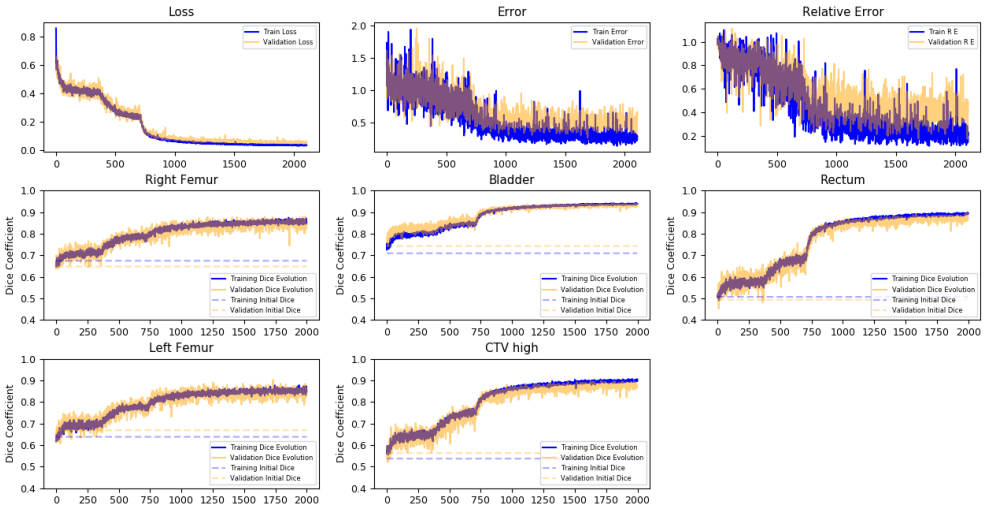


Figure 3.6: Training and validation curves of the MR-CT model trained on full image resolution. The training was monitored with the loss, euclidean error between the predicted and ground truth deformation fields, relative error to the average displacement, and Dice scores per organ. Blue represents the metrics for training data, while orange represents the metrics for validation data. The dashed lines depict the initial Dice scores.

To test the generalization capability in the multi-modality case two experiments were designed. The first one used the Iridium dataset by synthetically deforming the reference image and obtaining a ground truth deformed segmentation mask. The second experiment evaluated the capability of the models to generalize when facing real MR and CT images using the Gold Atlas dataset [25]. Both examples were benchmarked by comparing its performance with the one of ANACONDA algorithm.

3.2.2.1 Test on Iridium dataset

This experiment uses the test data samples from the Iridium dataset. This experiment was designed to test the generalization capability of the MR-CT models when facing new synthetic data. The resulting Dice coefficient scores per ROI can be found in Table 3.4 and the comparison of its Dice score distributions in Figure 3.7. Examples of resulting deformed images can be found in Figure 3.8.

The previous results show that the deep learning-based models learn to register the images improving the Dice scores after applying the predicted deformation

Model	Metric	R Femur	Bladder	Rectum	L Femur	Prostate
Full res	Dice	0.83	0.87	0.76	0.78	0.76
	Std	0.07	0.06	0.1	0.07	0.09
025	Dice	0.82	0.89	0.82	0.82	0.79
	Std	0.06	0.04	0.06	0.06	0.10
025 complete	Dice	0.93	0.93	0.89	0.87	0.87
	Std	0.01	0.02	0.03	0.04	0.06
RS	Dice	0.65	0.77	0.67	0.7	0.62
	Std	0.16	0.08	0.16	0.10	0.14
Benchmark scores	Dice	0.94	0.93	0.90	0.94	0.82

Table 3.4: MR-CT experiments' average Dice scores and standard deviation per model per organ on the Iridium dataset. The benchmark scores are from a deep learning segmentation model trained on the same data set [7].

fields. Yet, compared to the benchmark Dice scores, ours are still a little bit lower. From the comparison plot, it can be observed that the 025 complete model is the one that performs better. Nevertheless, it is still needed to test its performance on real images.

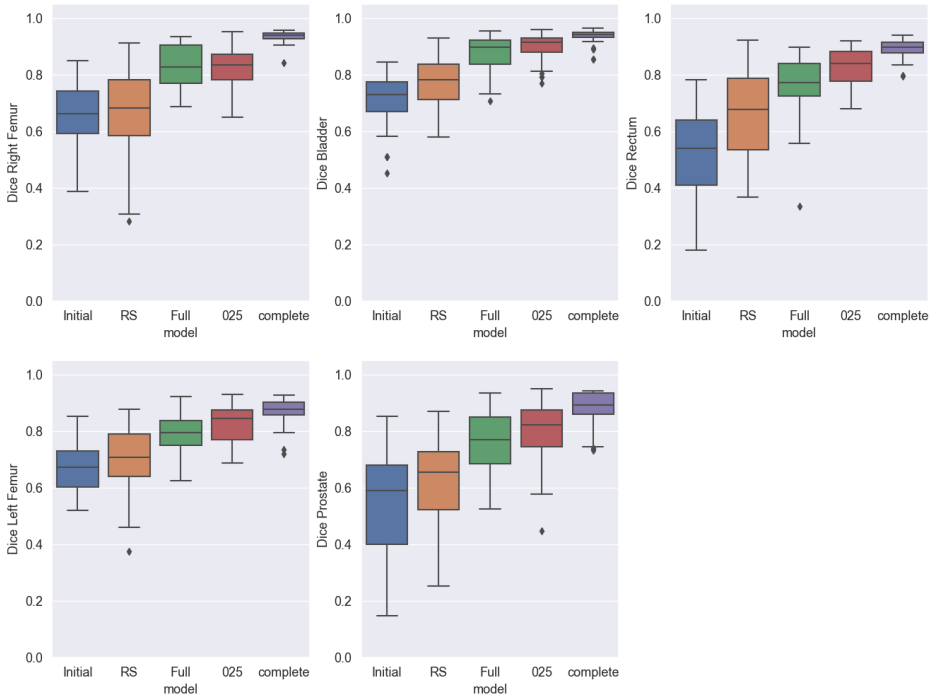
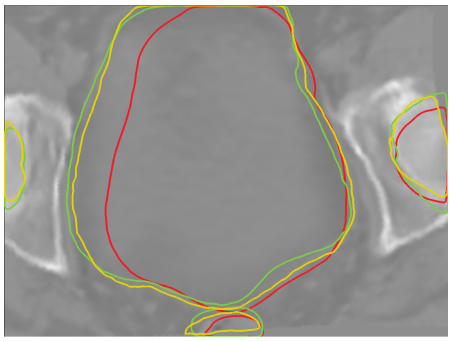


Figure 3.7: MR-CT model comparison of Dice coefficient scores per organ on Iridium synthetic data.

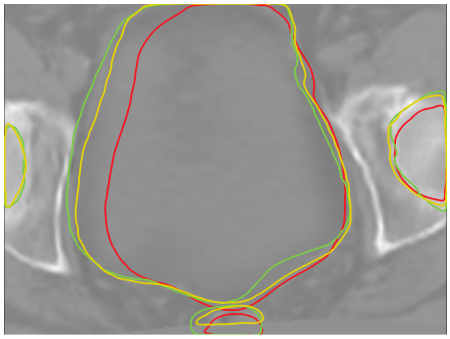
3.2.2.2 Test on Gold Atlas dataset

In this experiment, the capability of the multi-modal models to generalize on real T2 MR and CT data was tested using the Gold Atlas dataset. In the same way, as in previous tests, the performance of the models was compared to the results of applying ANACONDA algorithm on the same data samples. The resulting Dice coefficient distributions are presented in Figure 3.9 and the examples of deformed test images can be found in Figure 3.10. The resulting Dice coefficient scores per ROI can be found in Table 3.5.

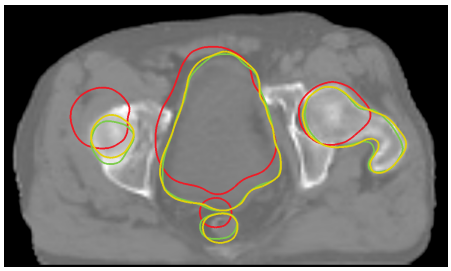
As it was expected, the performance of the model on real images was quite lower than when facing synthetic images. More precisely, the 025 complete model is not able to improve the Dice scores in all the ROIs. However, the Full and 025 models overperform the analytic algorithm on all the ROIs.



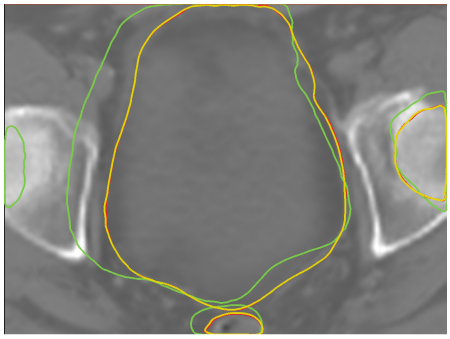
(a) Deformation result full resolution model



(b) Deformation result 0.25cm image resolution model.



(c) Deformation result 0.25cm image resolution model trained on non-cropped images.



(d) Deformation result of applying ANACONDA algorithm.

Figure 3.8: Comparison of the final registrations obtained by the different MR-CT models on the same test patient of the Iridium dataset. The background image is the deformed target image. In red the initial non-deformed masks, in green the reference masks, and in yellow the predicted deformed masks. Subfigure 3.8c has a different deformation because the deformations on the complete image sets were created separately from the ones for the cropped image set. Yet, they all have the same distribution.

3.2.3 Runtime Analysis

Another aspect in which deep learning-based methods can provide a potential improvement compared to its analytical counterparts is the execution time. It is claimed that once a model is trained, its prediction time is way lower than computing a registration by means of optimization algorithms. Therefore, the runtimes at inference for the different models are presented in Table 3.6.
After comparing the runtimes of the deep learning models at inference to the ones of ANACONDA algorithm it can be concluded that the deep learning-

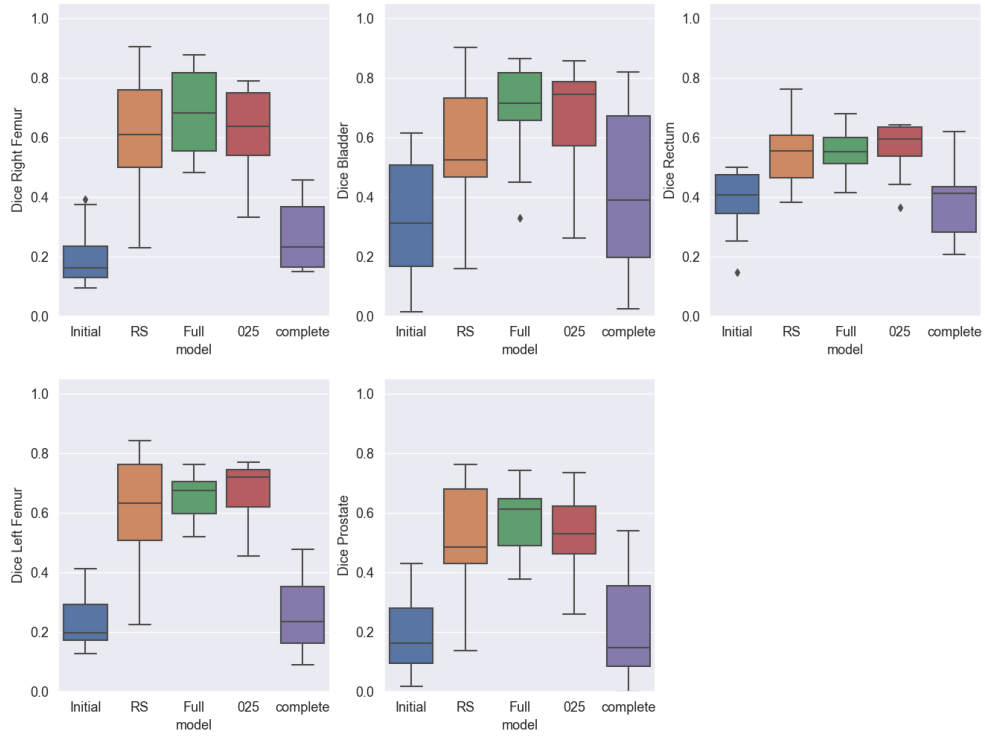
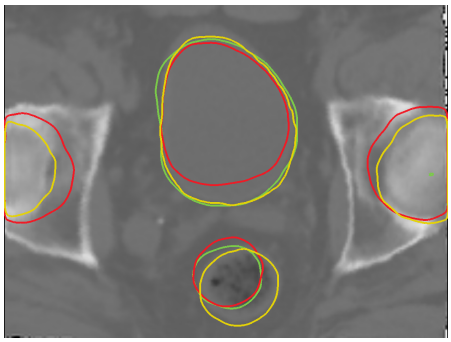


Figure 3.9: MR-CT models' comparison of Dice coefficient score per organ on the Gold Atlas dataset.

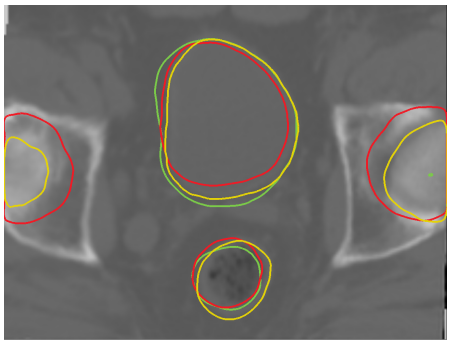
Model	Metric	R Femur	Bladder	Rectum	L Femur	Prostate
Full res	Dice	0.68	0.69	0.55	0.65	0.57
	Std	0.14	0.16	0.07	0.07	0.10
025	Dice	0.62	0.66	0.56	0.66	0.53
	Std	0.14	0.18	0.08	0.09	0.13
025 complete	Dice	0.28	0.41	0.38	0.26	0.21
	Std	0.11	0.27	0.12	0.12	0.17
RS	Dice	0.63	0.51	0.54	0.62	0.48
	Std	0.20	0.27	0.11	0.18	0.18
Benchmark scores	Dice	0.94	0.93	0.90	0.94	0.82

Table 3.5: MR-CT experiments' average Dice scores and standard deviation per model per organ on the Gold Atlas dataset. The benchmark scores are from a deep learning segmentation model trained on the same data set [7].

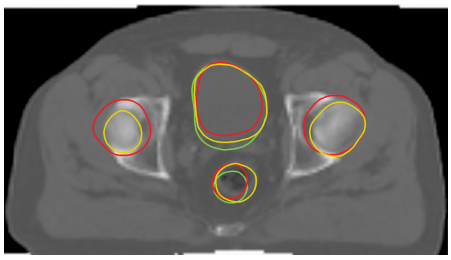
based models are much quicker, having a nearly constant runtime regardless of the difficulty of the registration. Additionally, their runtime is proportional



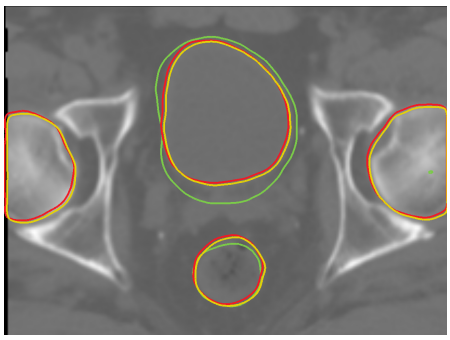
(a) Deformation result full resolution model



(b) Deformation result 0.25cm image resolution model.



(c) Deformation result 0.25cm image resolution model trained on non-cropped images.



(d) Deformation result of applying ANACONDA algorithm.

Figure 3.10: Comparison of the final registrations obtained by the different MR-CT models on the same test patient of the Gold Atlas dataset. The background image is the deformed target image. In red the initial non-deformed masks, in green the Reference masks, and in yellow the predicted deformed masks.

to the input image resolution. On the other hand, ANACONDA's runtimes are higher and very dependant on the type, difficulty of the registration task and the algorithm parameters.

Model	Average Time (s)
CT-CT 0107	2.74
CT-CT 0144	2.74
CT-CT 025	2.07
CT-CT 05	2.00
MR-CT Full res	3.37
MR-CT 025	2.47
MR-CT complete	2.4
RS CT-CT	8.69
RS MR-CT	5.13
RS MR-CT complete	5.30

Table 3.6: Average runtimes at inference for the deep learning based models compared to the ones for ANACONDA algorithm (RS).

Chapter 4

Discussion

In this project, a supervised registration algorithm based on a U-net like convolutional neural network has been proposed. Our approach follows the steps of [12], where ground truth data is generated to train a neural network in a supervised manner. Other approaches like the one in [8] proposed to train a deep similarity metric to be used in an optimization-based registration workflow. This approach was not suitable for our problem as one of the main motivations was to avoid the time cost of optimization algorithms. On the other hand, in [14], a method similar to the one implemented in the project was published, a neural network was trained to predict a deformation field but the loss function was defined as the similarity between the wrapped target image and the reference one. This semi-supervised approach does not suit the multi-modal registration task, being the method presented in [12] the most adequate for our purpose.

Following the procedure in [12], the lack of available ground truth registration data was overcome by generating synthetic deformation fields. For the MR-CT case, synthetic CT images were generated using a CycleGAN network from T2-MR images. This allowed to have perfectly aligned input image pairs. Furthermore, in most models, the images were cropped to a smaller patch to make sure that the network parameters would fit the memory requirements and focus on the important regions of interest: femoral heads, bladder, rectum, and prostate.

In the first place, the architecture presented in [12] was tailored to solve CT-CT registration of lung images, therefore a grid search was performed to optimize the baseline model to solve the task of registering CT-CT male pelvic

images. With the resulting optimization, only a small improvement of 0.03 was obtained compared to the architecture used in [12], suggesting that future improvements of this approach will not come from further fine-tuning of the U-net architecture. To see the effect of the image resolution during the network's learning process, four models were trained on different image resolutions, both an- and isotropic. Despite the results showed that the deep learning models trained on higher resolution images had a better performance than those trained on lower image resolutions, all models obtained Dice scores higher than 0.82 on synthetic test images. The average Dice score for the model trained on isotropic higher image resolution (1.44mm) was 0.87 and the average Dice score for the model trained on lower image resolution was 0.84. This is a surprisingly small difference considering that the high-resolution model uses 42 times as many voxels. Therefore, the resolution of the training images did not have a major impact on the models' performance.

All CT-CT models performed well and showed a similar performance to the clinically validated segmentation model (Table 3.3) trained on the same data. This comparison is interesting since the problem of image registration is strongly related to the problem of image segmentation, and a reasonable guess is that an optimally trained registration model would perform equal to or better than the segmentation. However, this result is valid only for synthetic registrations and the performance on real data has not been tested.

Secondly, three MR-CT models were trained with different input information: cropped images with full image resolution, 0.25 resolution cropped images, and 0.25 resolution complete images. The results showed that the two first models outperformed the optimization-based algorithm when being tested on both synthetic and real images. However, the obtained Dice scores from the synthetic images were slightly lower compared to the CT-CT model. This indicates, as expected, that multi-modal image registration is more difficult to learn than mono-modal. The difference was quite small though, 0.88 for the best CT-CT model and 0.82 for the MR-CT case. This should be compared with the much larger decrease in performance of the analytic algorithm ANACONDA, with average Dice scores of 0.91 for CT-CT registration and 0.68 for MR-CT registration. This demonstrates that deep learning is well suited for multi-modal registration and does not suffer from the same problems as analytical methods that rely on mutual information as similarity metric. In addition, the performance of the deep learning models drops when tested on real data. Yet, the models trained on cropped images still outperform the state-of-the-art

analytical algorithms, here represented by the ANACONDA implementation in RayStation. The average Dice scores of the best performing deep learning model and ANACONDA were 0.63 and 0.55 respectively. Even though it is questionable if the model performs well enough to be useful in practice, we consider this a promising result and a stepping stone for developing highly accurate multi-modal registration models.

After analyzing the results it is clear that the proposed method has certain limitations. On one hand, the choice of parameters and method used to create ground truth deformation fields needs to be carefully chosen given that it needs to be as realistic as possible. The pelvic region is a part of the body that suffers deformations that are difficult to simulate artificially. It has organs such as the bladder and the rectum that can change of shape and filling at any time, while the hip and femur bones may only suffer from translations. When designing the synthetic deformations for this project, it was assumed that deforming image regions that would not be deformed in a real case would not affect the registration capability of the network on real data. This assumption was made to simplify the difficulty of the initial problem. Additionally, in most cases, the images were cropped to ensure the model parameters fitted the available memory space, and to focus on relevant regions such as the femoral heads, the bladder, the prostate and the rectum. As each patient is different and, therefore, each image is different, the cropping of the images caused the femoral heads to appear cropped in some images, affecting the network performance in these areas.

Chapter 5

Conclusions and Future Work

A supervised architecture for deformable registration of pelvic images has been presented. The architecture is based on a U-net like neural network trained to estimate the deformation vector field to register a target CT image onto a T2-MR reference one. In this project, the architecture has been trained on CT-CT images for mono-modal image registration and on MR-CT images for the multi-modality case, using synthetic deformations to generate the ground truth deformation vector fields. Combined with initial rigid-body registration, the model could accurately register the synthetic test data for both the mono-modal and multi-modal case. For the multi-modal case, the network was able to improve the Dice scores when compared to the initial ones on real data and overperform the optimization-based benchmark method. This suggests a great potential of using deep learning for deformable image registration bringing advantages such as a lower time and cost of image registration in the clinics.

Given the 5 months scope of this project several aspects were kept for future research. The first one is to investigate a way of generating realistic synthetic deformations. In this project, a combination of a fine and a coarse grids were used to generate a random deformation field, yet, due to its randomness, it may not be realistic. This may have led to a decrease in the performance of the model when facing real images. In terms of the neural network architecture, a natural future work step is to implement a stacked architecture where the intermediate network's output is the input for the next network. Additionally, it would be interesting to investigate into more detail the effect of scheduling data presentation to the network and how would that affect its learning as presented in [17]. All in all, this was only a preliminary study about the viability of

using deep learning for deformable image registration, and there is still plenty to explore.

Bibliography

[1] Mean absolute error. https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/loss-functions/mean-absolute-error. Accessed: 2020-03-11.

[2] Mean squared error. https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/loss-functions/mean-squared-error. Accessed: 2020-03-11.

[3] Bladder cancer: Statistics. https://www.cancer.net/cancer-types/bladder-cancer/statistics, 05-2019. Accessed: 2020-01-29.

[4] Metrics to evaluate your semantic segmentation model. https://towardsdatascience.com/metrics-to-evaluate-your-semantic-segmentation-model-6bcb99639aa2, 09-2019. Accessed: 2020-03-11.

[5] Prostate cancer: Statistics. https://www.cancer.net/cancer-types/prostate-cancer/statistics, 11-2019. Accessed: 2020-01-29.

[6] C. Davatzikos A. Sotiras and N. Paragios. Deformable medical image registration: A survey. *IEEE Transactions on Medical Imaging*, pages 1153–1190, July 2013.

[7] Raysearch Laboratories AB. Validation report pelvic male segmentation iridium.

[8] Alireza Mehrtash Steve Pieper Clare M. Tempany Tina Kapur Parvin Mousavi Alireza Sedghi, Jie Luo and William M. Wells III. Semi-supervised deep metrics for image registration. Apr 2018.

[9] Katherine Anne Bachman. Mutual information-based registration of digitally reconstructed radiographs and electronic portal images. 2006.

[10] ATAM P. DHAWAN. *MEDICAL IMAGE ANALYSIS*, chapter Chapter 12 - Image Registration. JOHN WILEY SONS, INC, 2011.

[11] Dan Ruan Daniel O'Connor Minsong Cao Elizabeth M. McKenzie, Anand Santhanam and Ke Sheng. Multimodality image registration in the head-and-neck using a deep learning-derived synthetic ct as a bridge. Dec 2019.

[12] K. A. J. Eppenhof and J. P. W. Pluim. Pulmonary ct registration through supervised learning with convolutional neural networks. *IEEE Transactions on Medical Imaging*, 38(5):1097–1105, 2019.

[13] Yabo Fu, Yang Lei, Tonghe Wang, Walter J. Curran, Tian Liu, and Xiaofeng Yang. Deep learning in medical image registration: A review, 2019.

[14] M. R. Sabuncu A. V. Dalca G. Balakrishnan, A. Zhao and J. Guttag. An unsupervised learning model for deformable medical image registration. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9252–9260, Dec 2018.

[15] Niko Papanikolaou Geoffrey Clarke. Mri for diagnosis and mri for diagnosis and treatment of cancer treatment. https://www.aapm.org/meetings/amos2/pdf/34-8205-79886-720.pdf. Accessed: 2020-05-08.

[16] Kruger U. Yan P Haskins, G. Deep learning in medical image registration: a survey. *Machine Vision and Applications*, 31, 2020.

[17] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. *CoRR*, abs/1612.01925, 2016.

[18] P. Isola J. Zhu, T. Park and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017.

[19] Nagarajan Kandasamy James Shackleford and Gregory Sharp. *High Performance Deformable Image Registration Algorithms for Manycore Processors*, chapter Chapter 1 - Introduction. Morgan Kaufmann, 2014.

[20] Swamy Laxminarayan Jasjit S.Suri, David L.Wilson. *Handbook of Biomedical Image Analysis; Volume III: Registration Models*, chapter 1.2.3 Nature of Transformation. Kluwer Academic/Plenum Publishers, 2005.

[21] Sherif Abdulatif Thomas Kustner Sergios Gatidis Bin Yang Karim Armanious, Chenming Jiang. Unsupervised medical image translation using cycle-medgan. Mar 2019.

[22] Artan Kaso. Computation of the normalized cross-correlation by fast fourier transform. *Plos One*, Set 2018.

[23] Diana Mateus Nassir Navab Nikos Komodakis Martin Simonovsky, Benjamín Gutiérrez-Becker. A deep metric for multimodal registration. *19th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2016)*, pages 10–18, Oct 2016.

[24] H.B. Mitchell. *Image Fusion: Theories, Techniques and Applications*, chapter Image Similarity Metrics. Springer International Publishing.

[25] Tufve Nyholm, Stina Svensson, Sebastian Andersson, Joakim Jonsson, Maja Sohlin, Christian Gustafsson, Elisabeth Kjellén, Karin Söderström, Per Albertsson, Lennart Blomqvist, Björn Zackrisson, Lars E. Olsson, and Adalsteinn Gunnlaugsson. Mr and ct data with multiobserver delineations of organs in the pelvic area—part of the gold atlas project. *Medical Physics*, 45(3):1295–1300, 2018.

[26] Jinyoung Kim Guillermo Sapiro Reuben R Shamir, Yuval Duchin and Noam Harel. Continuous dice coefficient: a method for evaluating probabilistic segmentations. Apr 2018.

[27] Brox T Ronneberger O., Fischer P. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015.*, 9351, May 2015.

[28] William Rucklidge, editor. *The Hausdorff distance*, pages 27–42. Springer Berlin Heidelberg, Berlin, Heidelberg, 1996.

[29] Amod Jog Jerry L. Prince Snehashis Roy, Aaron Carass and Junghoon Lee. Mr to ct registration of brains using image synthesis. *Medical Imaging 2014: Image Processing*, March 2014.

[30] Dr. S. Arivazhagan V.R.S Mani. Survey of medical image registration. *Journal of Biomedical Engineering and Technology*, pages 8–25, 2013.

[31] Ola Weistrand and Stina Svensson. The anaconda algorithm for deformable image registration in radiotherapy. *Medical Physics*, 42(1):40–53, 2015.

[32] Yanwei Pang Eric Granger Xiaoyi Feng Xiaoyue Jiang, Abdenour Hadid. *Deep Learning in Object Detection and Recognition*, chapter Chapter 1 - 1 Brief Introduction. 3, 2019.

[33] Tonghe Wang Walter J. Curran Tian Liu Yabo Fu, Yang Lei and Xiaofeng Yang. Deep learning in medical image registration: A review. Dec 2019.

[34] Wiro J. Niessen Yuanyuan Sun, Adriaan Moelker and Theo van Walsum. *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, chapter Towards robust ct-ultrasound registration using deep learning methods. Springer International Publishing.

Appendix A

Background

A.1 Image Registration

Image registration is defined in [16] as the act of mapping the content of two images to the same coordinate system. The difference between the images may be due to its acquisition at different times, from different angles or modalities (multi-modality). Within the pair of images to be registered one will be the reference image, R, and the other the one to be transformed, target image$M$. Yet, the original images must contain information about the same object or structure. Both images are related by a transformation. The goal of image registration is to estimate the transformation that optimizes a cost function of the form of Equation A.1 in order to obtain a more accurate lineup mapping between the reference and target image.

$$S(R, MoT) + \lambda(T) \tag{A.1}$$

Equation A.1 presents an objective function where S denotes the similarity metric which quantifies the level of alignment between the reference and target images. T denotes the transformation applied to the target image. λ represents the regularization term added to the transformation to encourage specific properties in the solution. Both reference and target images are defined in the image domain Ω as well as T which maps homologous locations from the reference image to the target image [6]. The basic image registration flowchart consists of four basic steps. The first one is to choose a random set of starting parameters. Secondly, apply the transformation based on the previous parameters to the target image by means of an interpolator to lineup the reference image to the target one. Thirdly, evaluate the cost function based on the chosen similarity metric between the target and reference image. Next, the convergence

criteria must be checked. If the convergence criteria has been met, the registration procedure is finished. Otherwise, the optimizer should find a new set of parameters for the transformation and iterate throughout the process again. The overall flowchart of the process is depicted in Figure A.1.

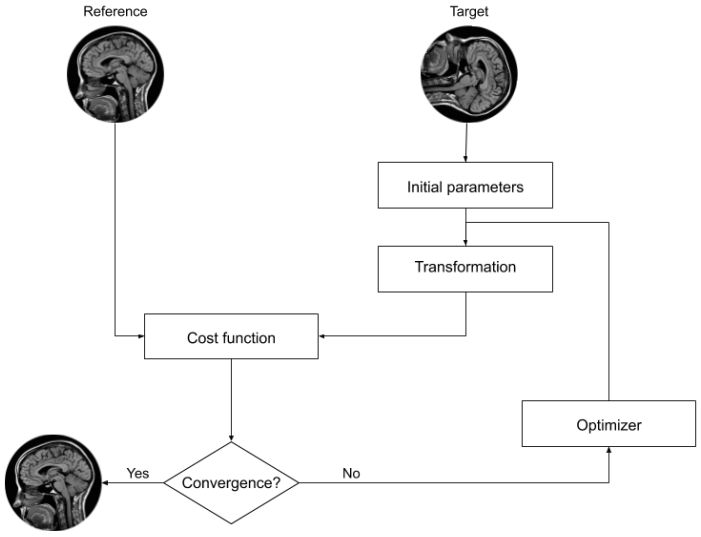


Figure A.1: Image registration flowchart.

A.1.1 Nature of Transformation

In order to select properly the registration method to be used, it is necessary to consider the type of deformation that the images have faced. In the following subsections the four main types of transformations will be explained in further detail.

A.1.1.1 Rigid-Body Transformation

Rigid-Body transformations are based on translation and rotation operations to the original images. This means that two images of equal dimensions are registered by means of a pixel-wise transformation consistent along the image space. Equation A.2 shows the application of rotation R and translation t to the original image x to obtain the registered result x' [10].

$$x' = Rx + t \tag{A.2}$$

A.1.1.2 Affine Transformation

Affine transformations can be considered a type of Rigid-Body transformations as it also includes translation and rotation operations in addition to scaling and shearing. This kind of transformation is used to register pairs of images in different scales, preserving parallel lines but not their lengths or angles. In this case, a scaling and shearing factor is added to each image dimension increasing the degrees of freedom of the transformation [30]. Affine transformations can be expressed following Equation A.3, where A is the affine matrix which includes rotation, translation, scaling and shearing transformation parameters [10].

$$x' = Ax \tag{A.3}$$

A.1.1.3 Projective Transformation

Projective transformations are used when the image appears tilted. This kind of transformation preserves straight lines but parallel ones converge towards a vanishing point. Mapping in this way parallel lines from the target image to the reference one. This transformation is sometimes used as a "constrained elastic" transformation when the optimizer is unable to find a solution for the elastic registration [30].

A.1.1.4 Non-Rigid-Body Transformation

Contrarily to the previously presented transformation types, non-rigid transformations create a mapping between pixels through nonlinear dense transformations or spatially varying deformation fields [6]. This means that non-rigid transformations are capable of expressing nonlinear relations, being able to map lines on to curves [20]. Therefore, they are also called elastic or deformable registrations. When performing deformable registrations, a dense, non-linear correspondence is established between a pair of n-dimensional volumes. Most image registration methods solve this problem by optimizing a similarity function for each voxel pair, which aligns voxels having a similar appearance and enforces smoothing constraints when computing the registration mapping [14]. Most of the deformable registration methods presented in the literature follow a two-step workflow performing, in the first place, rigid registration followed by a deformable registration.

A.1.2 Similarity Metrics

One of the most important parts of the image registration pipeline is to determine an appropriate similarity metric to asses the quality of the registration. In this section, the most relevant similarity metrics related to deformable image registration are going to be described. In the following subsections the formulas of the similarity metrics are also presented, in these A is the reference image, B is the wrapped target image, a is a pixel of A and b is a pixel of B. Finally, N represents the total number of pixels in the image.

A.1.2.1 Dice Coefficient

Dice Coefficient is a very widely used similarity metric for image segmentation. This measures the overlap between the calculated segmentation and the ground truth one. Dice Coefficient is calculated as the double of the intersection between the segmentations divided by the sum of the total number of pixels of both masks [4]. Its values range from 0 to 1, being 1 a perfectly overlapping segmentation. Its formula can be found in Equation A.4.

$$Dice = \frac{2 \times (A \cap B)}{A + B} \tag{A.4}$$

A.1.2.2 Jaccard Coefficient

This similarity metric measures the overlap between the predicted segmentation image over the ground truth mask divided by the area of union between the computed and the ground truth segmentations. The jaccard coefficient values range from 0 to 1, being 1 a perfectly overlapping segmentation [4]. The jaccard coefficient formula is stated in Equation A.5.

$$JC = \frac{A \cap B}{A \cup B} \tag{A.5}$$

A.1.2.3 Normalized Cross Correlation (NCC)

Normalized Cross Correlation measures the amount of correlation between the two images. Its normalized version is used to avoid the dependence of the covariance on the amplitude of the compared images [22]. Its mathematical expression can be found in Equation A.6.

$$NCC = \frac{A \times B}{|A|\,|B|} \tag{A.6}$$

A.1.2.4 Mutual Information (MI)

This metric estimates the joint probability that a pixel in the reference image has the same intensity value as the same pixel in the target image. The metric is described in the Equation A.7. Where $p_A(a)$ is the probability a pixel in A has an intensity value a, $p_B(b)$ is the probability a pixel in B has an intensity value b, and $p_{AB}(a,b)$ is the joint probability a pixel in A has an intensity value a that is the same as b in image B [24]. Its formula is stated in Equation A.7.

$$MI(A,B) = \int\int p_{AB}(a,b) log_2 \frac{p_{AB}(a,b)}{p_A(a)p_B(b)} dxdy \tag{A.7}$$

A.1.2.5 Normalized Mutual Information (NMI)

Normalized Mutual Information is introduced as an improvement to MI. The problem of MI is that the integral is taken over the pixels which are common to both A and B. Therefore, if the common number of pixels between A and B changes the MI(A,B) metric will also change. Even if these changes may be small they may lead to inaccuracies in the registration algorithm. To avoid that, normalized mutual information is used in place of MI [24]. The equation of NMI is described in Equation A.8.

$$NMI(A,B) = \begin{cases} \frac{MI(A,B)}{H(A)+H(B)} \\ \frac{MI(A,B)}{min(H(A)+H(B))} \\ \frac{MI(A,B)}{H(A,B)} \\ \frac{MI(A,B)}{\sqrt{H(A)H(B)}} \end{cases} \tag{A.8}$$

where the entropies are calculated as follows:

$$\begin{aligned} H(A) &= -\int\int p_A(a) log_2 p_A(a) dxdy \\ H(B) &= -\int\int p_B(b) log_2 p_B(b) dxdy \\ H(A,B) &= -\int\int p_{AB}(a,b) log_2 p_{AB}(a,b) dxdy \end{aligned} \tag{A.9}$$

A.1.2.6 Mean Squared Error (MSE)

Mean Squared Error is one of the most widely used as a loss function metric for regression problems. However, it can also be used to calculate mean squared

differences between the wrapped target image and the reference one [2]. Its formula can be found in Equation A.10.

$$MSE = \frac{1}{N}\sum^{N}(a-b)^2 \tag{A.10}$$

A.1.2.7 Mean Absolute Error (MAE)

Mean Absolute Error is usually applied as a loss function metric for regression problems. However, it can also be used to calculate the differences between the wrapped target image and the reference one [1]. Its formula can be found in Equation A.11.

$$MAE = \frac{1}{N}\sum^{N}|a-b|^2 \tag{A.11}$$

A.1.2.8 Hausdorff Distance

Hausdorff Distance is a distance metric used to compare two objects and find the maximum distance between them [28]. Its formula is described in Equation A.12.

$$H(A,B) = \max(h(A,B), h(B,A)) \tag{A.12}$$

were

$$h(A,B) = \sup_{a\in A}\inf_{b\in B}||a-b|| \tag{A.13}$$

A.2 Deep Learning for Image Registration

Deep Learning (DL) methods are defined as a class of machine learning algorithms that make use of many layers of nonlinear processing units to create representations and transformations to map the input values to the output ones. The architecture uses the output from the previous layers as input, creating a hierarchical representation that can be obtained by different levels of abstraction [32]. DL algorithms can be trained in a supervised or unsupervised fashion depending on its application and availability of data. Supervised methods involve the designation of a ground truth output for the neural network. Whereas unsupervised methods draw inferences given the probability distribution of the given data without any defined ground truth label. This field has experienced a rapid development thanks to the recent availability of data, powerful electronic

devices that ease consuming computations and the development of novel algorithms.

Given the time, computational, and economic constraints of traditional image registration methods, automatic deep learning methods have been applied to find solutions to the image registration task. Image registration is highly needed in many fields, but one of the most important applications is to register medical images. Therefore, from now onwards this chapter is going to focus on medical image registration methods. In an early stage of DL-based registration methods, DL was applied to medical images to augment the performance of iterative, intensity-based registration pipelines. Later, reinforcement learning approaches were researched to perform image registration. Given the demand for faster registration methods, deep-learning-based one-step transformation estimation techniques were developed. Nevertheless, the unavailability to obtain ground truth data has motivated the development of unsupervised algorithms for one-step transformation estimation [16]. The field of DL based image registration is evolving at a quick peace given that there is a need for quick and high-quality image registration. Figure A.2 shows a visual representation of the rapid evolution of the field extracted from [16]. These different methods are going to be explained in the following section.

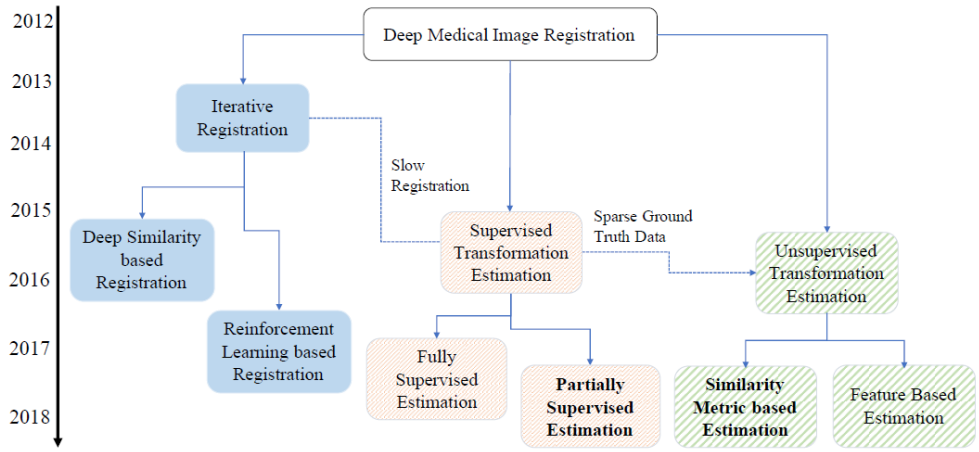


Figure A.2: Overview of the evolution of deep learning based medical image registration methods from [16].

A.2.1 Methods

In this section, the three main approaches for medical image registration that are present in the literature are going to be explained in further detail.

A.2.1.1 Deep Iterative Registration

This group of methods is characterized by repeating the same operation along several iterations to obtain a better result. In this category two sub-methods can be found [16]:

- *Deep Similarity-based Registration*: This set of methods use deep learning to learn a similarity metric to be inserted into the basic intensity-based registration framework with a defined interpolator, transformation model and optimization algorithm. The results stated in the literature suggest that deep learning is capable of learning a similarity metric for the multi-modal registration case.

- *Reinforcement Learning based Registration*: These methods use a trained agent to perform the registration task as opposed to a defined optimization algorithm. These methods normally involve a rigid transformation model but it is still possible to use a deformable one.

A.2.1.2 Supervised Transformation Estimation

This method significantly speeds up the registration process when compared to the aforementioned given that it uses a neural network for the registration task and not an iterative optimizer. Its main goal is to perform a single-step method to estimate the transformation parameters directly. This method can be divided into two different approaches:

- *Fully Supervised Transformation Estimation*: This approach uses full supervision and ground truth labels for the one-step transformation estimation task. The advantage of this kind of methods is that the performance of deformable registration does not add extra constraints when compared to rigid registration methods. Yet, it also has its drawbacks. When using this method, the quality of the registrations depend directly on the quality of the ground truth registrations. This data can be created synthetically however it is important to ensure that it is similar to the real case clinical data.

- *Dual/Weakly Supervised Transformation Estimation*: This method makes use of both ground truth data and a similarity metric to quantify the similarity to train the model. Contrarily, weak supervision refers to the usage of overlapped segmentations of corresponding anatomical structures to design the loss function.

A.2.1.3 Unsupervised Transformation Estimation

This method issued from the difficulty of acquiring reliable ground truth labels for the registration task. An architecture that has enabled the development of these techniques is the spatial transformer network (STN). The STN has been used to perform deformations associated with the registration applications. This approach can be divided into two major methods:

- *Similarity Metric-Based Unsupervised Transformation Estimation*: This set of methods use a similarity metric with common regularization strategies to define the loss function. This approach has received a lot of attention from the research community as it avoids the need of expert labels and the performance of the model is not dependent on the expertise of the practitioner. Given that it is still difficult to quantify the image similarity of multi-modal registration applications, this approach is reserved for the mono-modal case.

- *Feature-Based Unsupervised Transformation Estimation*: This set of methods does not require ground truth data as it learns the representation of the features from the data and uses it to train the neural network.

A.2.1.4 GAN-based methods

Generative Adversarial Networks were first introduced by [13] as a framework for estimating generative models via an adversarial process by training two models simultaneously: a generative model which learns the distribution of the data, and a discriminative model in charge of estimating the probability that a sample comes from the dataset rather than being generated by the generative model. The aim of the network for the generative model is to maximize the probability that the discriminator is wrong. This architecture has been used for medical image registration for two main reasons. In the first place, it has been used as a regularization method of the predicted transformation, preventing this way unrealistic transformations by promoting smoothness, anti-folding, and inverse consistency constraint. GANs have also been used for multi-modal

image registration to transform the multi-modality problem to mono-modality by mapping images from one domain to another [33].

A.2.1.5 Summary

Different approaches have been taken to tackle the problem of image registration during the research of deep-learning-based methods. Firstly, Deep Iterative Registration methods were used. These, estimate either a similarity metric to perform image registration or train an agent to develop the registration task. The problem that this approach presents is that as the registration is performed throughout a number of iterations it is very time-consuming. To continue, Supervised Transformation Estimation methods use ground truth data in order to help the neural network to learn the wanted transformation. The advantage of this set of methods is that they can learn how to perform rigid-body and deformable registration without adding constraints to the network. Yet, their performance depends on the quality of the ground truth data. By exploring Unsupervised Transformation Estimation algorithms, new methods that needed few ground truth data or no data at all are introduced. This eliminates the dependency on ground truth data to train the model. Finally, GAN-based methods have helped to regularize the predicted transformation preventing folding and other unrealistic changes as well as to convert multi-modality problems to mono-modality. The method that is going to be used in the project is *Fully Supervised Transformation Estimation* as the goal is to solve the multi-modality problem having ground truth data. In Table A.1 a summary of the different presented methods can be found.

Deep Learning Methods Summary				
Method	Need Ground Truth	Estimate Similarity	Estimate Transformation	Time
Deep Iterative Registration	Yes/No	Yes	No	High
Supervised Transformation Estimation	Yes	No	Yes	Low
Unsupervised Transformation Estimation	No/Sparse	Yes	Yes	Low
GAN-based methods	Yes/No	No	Yes	Low

Table A.1: Deep learning based methods for image registration summary.

A.2.2 Important Architectures

In this section, two neural network architectures are going to be presented: CycleGAN and U-Net. Both architectures have been used in the field of medical image registration and introduced major improvements to the field.

A.2.2.1 CycleGAN Architecture

The CycleGAN architecture was first presented in [18] as a framework for learning to translate an image from a source domain X to a target domain Y when no paired ground truth images are available. The main goal of the network is to learn the mapping function G so that $G : X \rightarrow Y$ in a way that the distribution of $G(X)$ is indistinguishable from the distribution of Y using an adversarial loss. In order to constrain the previous mapping, the loss is coupled with an inverse mapping $F : Y \rightarrow X$ to introduce the cycle consistency loss and enforce that $F(G(X)) \approx X$ and vice-versa. The main idea behind the cycle consistency losses is to capture the intuition that if one image is translated from one domain to the other and back to the first one, the result should be the same as the input image. This procedure is explained graphically in Figure A.3.

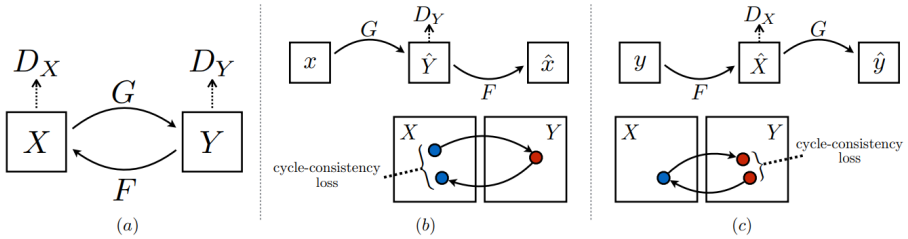


Figure A.3: Figure from [18] © 2017 IEEE. (a) CycleGAN model mapping functions G and F, and the associated discriminators D_Y and D_X. Where D_Y encourages G to translate X into outputs of the form of the domain Y, and the other way round for D_X and F. (b) and (c) show the forward and backward cycle-consistency loss respectively that capture the intuition that if we translate from one domain to the other and back again the output should be of the same domain as the first input.

In the field of healthcare images, the unavailability of coupled ground truth images is even a greater problem. Given this, CycleGAN architecture has been adapted to perform image to image translation between medical images of dif-

ferent modalities. This has been presented in [21], where CycleGAN architecture has been adapted to perform PET-CT translation and MR motion correction. The proposed neural network architecture is called Cycle-MedGAN and it is characterized by the inclusion of non-adversarial losses.

A.2.2.2 U-Net Architecture

One of the most popular DL architectures present in the literature to perform image registration is called U-net, introduced in [27]. This network was first introduced as a modification of a fully convolutional network to perform image segmentation, cell segmentation to be more precise. Its architecture is characterized by two main components. The first one is a contracting path that allows to extract relevant information from the input. This is followed by an expanding path which is symmetrical to the contracting one. This is important because features from the contracting path are combined with the up-sampled version of the expanding path allowing to localize high-resolution features. Moreover, the expanding path has a large number of feature channels allowing to propagate context information to higher resolution layers.
Another characteristic of U-net is that it relies on a very small amount of training samples, relying heavily on data augmentation. During training, elastic deformations are applied to the data samples, allowing the network to learn invariance to such deformations without having them in the original training set. In their case, this strategy is very useful for cell segmentation as realistic cell deformations can be simulated efficiently.
The architecture proposed in the original paper consists of a contracting and an expansive path. Firstly, the contractive path consists of two 3x3 unpadded convolutions followed by a rectified linear unit (ReLU), Equation A.14, and 2x2 max pooling with stride 2 for downsampling. At each downsampling step, the number of feature channels is doubled. On the other hand, a step of the expansive path consists of an upsampling of the feature map followed by a 2x2 up-convolution. The previous will halve the number of feature maps of the layer. However, this is later concatenated with the corresponding downsampled feature map from the contracting path and followed by two 3x3 convolutions that finish with a ReLU activation function. Finally, a 1x1 convolution is used at the final layer to map each feature vector component to the desired number of classes. The overall network architecture is represented in Figure A.4.

$$y = max(0, x) \tag{A.14}$$

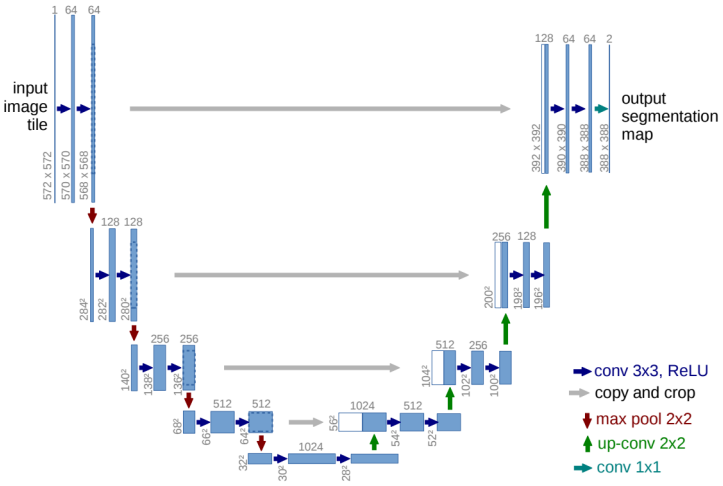


Figure A.4: U-net architecture from [27].

A.2.3 U-Net for Image Registration

The U-net architecture was originally created for semantic image segmentation. Yet, given its capability to extract context information, the network architecture has evolved to be used in different image analysis fields such as image registration. In [12], a modified version of U-net is presented to solve the problem of mono-modal deformable registration field estimation. In their work, four main modifications are introduced to the original network. The first one is to feed the network with two inputs: the target and the reference images. Secondly, the architecture is deepened one more level. Then, the activation function is changed from ReLU to Leaky ReLU, Equation A.15. And finally, the output convolutional layer of the network is changed to have three feature maps, one for each dimension of the displacement field to be predicted. To train the network, L_1 norm of the absolute error between the ground truth deformation field and the predicted one is used as the loss function.

$$y = 0.01x; \ when \ x < 0 \tag{A.15}$$

One of the main issues when performing image registration is the availability of data. It is very difficult to obtain registered ground truth images and in the case that these can be obtained they are very expensive. In [12], this issue is addressed by synthetically generating a deformation field and applying it to an image, this way a ground truth field is obtained as well as the two input images. Additionally, when dealing with clinical data it is often quite difficult to obtain a large extension of images to work within the first place. As it was done in

the original U-net publication, *Eppenhof K.A.J. et al.* use data augmentation to generate new data samples. The graphic representation of the proposed network architecture in [12] can be found in Figure A.5.

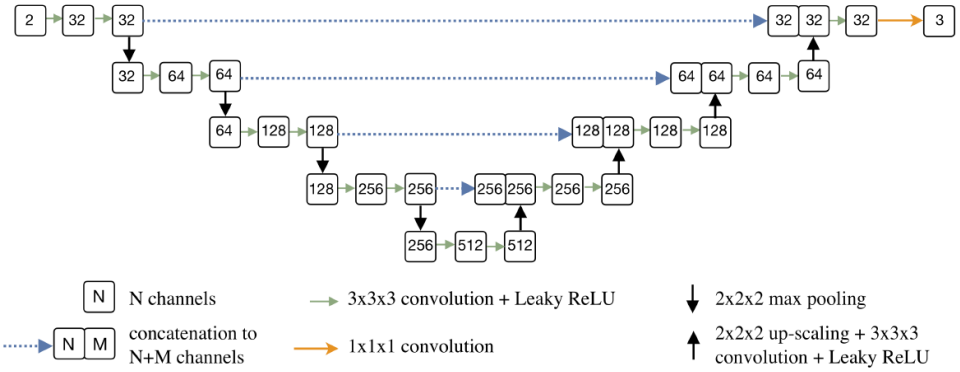


Figure A.5: Network architecture from [12] © 2018 IEEE.

Additionally, another U-net like architecture to approach the challenge of image registration is presented in [14]. In the publication, two architectures are proposed for solving the challenge of MRI elastic registration. These architectures are called VoxelMorph-1 and VoxelMorph-2, where the difference between them relies on the number of channels at the outer layers and the number of final convolutional layers. These architectures are shown in Figure A.6.
In addition, a different training approach is presented. Their network architecture receives the reference and target images as input and outputs a deformation field. Yet, this deformation field is applied to the target image, and the loss of the network is calculated as the cross-correlation between the transformed target image and the reference one.

A.2.4 The multi-modality problem

All the approaches presented in Section A.2.3 are used to face the problem of mono-modal image registration. This means that both reference and target images are of the same type, MR-MR, CT-CT, PET-PET, etc. When dealing with multi-modal image registration there are two main issues to overcome. The first one is to use an appropriate similarity measure. Different modality images have different pixel intensity scale and have different intensity values for the same representation of the feature. This causes some similarity metrics that measure the overlap or the similarity between pixel values to not be suit-

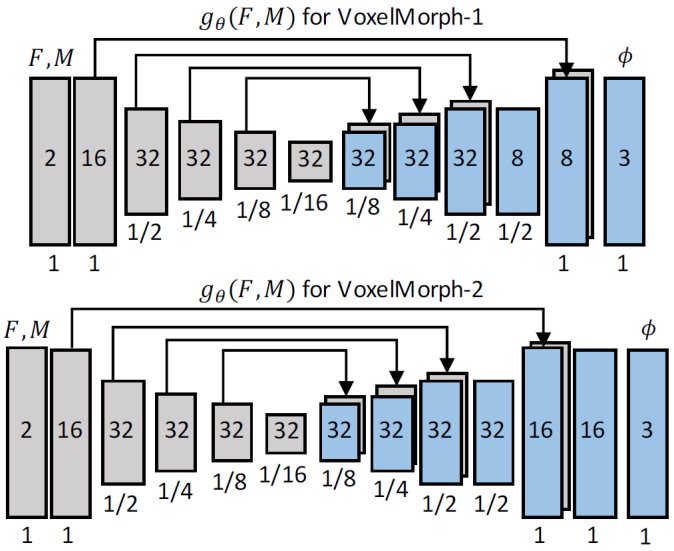


Figure A.6: Network architectures from [14], where $g_\theta(F, M)$ represents the transformation function to be applied on the target image © 2018 IEEE.

able for the deformable registration task. The standard metric for multi-modal registration is mutual information (MI) [23]. This method compares the statistical relationship among single pixels from the different image modalities. Although it has been widely used to solve the multi-modal registration problem, to simply measure pixel statistics may not be efficient enough to capture all the information of the images [8]. Moreover, MR modality intensity does not have an accepted calibrated intensity scale. This means that the overall image intensity distribution can vary from scan to scan. This can be a problem as MI depends on the joint histogram of the images [29]. Therefore, another approach proposed in the literature is to develop a neural network to learn a deep similarity metric to be used during the multi-modal image registration[8][23]. The second main issue to be faced when performing multi-modal image registration is the availability of data. As multi-modality implies that the images are acquired at different times and normally by different acquisition machines, its lineup is never guaranteed. Additionally, in order to have ground-truth data, registrations produced by experts are needed, making them expensive and very dependant on the expertise of the professional. To overcome this second problem synthetic images are produced. As a first step before building the registration network architecture, a CycleGAN network is built and trained to generate synthetically images from one domain to the other. An example of this method

can be found in [29] and [11] where synthetic CT images are generated from MR ones. Then, as it has also been introduced in Section A.2.2.2, synthetic deformation fields are applied to the input images to augment the available data. The risk of taking this approach is that the final model may not generalize when facing real data as presented in [34].

A.3 ANACONDA Deformable Image Registration

ANAtomically CONstrained Deformation Algorithm (ANACONDA) [31] is an analytic registration method that combines image information with anatomical information provided by contoured image sets. In this method, the objective function to minimize during optimization is a linear combination of nonlinear terms: image similarity, grid regularization, shape-based regularization, and a penalty term added when controlling structures are used to compute the registration. Each term has its own contribution to the registration task. The image similarity term measures the similarity between the reference and the target after applying the computed deformation. The grid regularization term keeps the deformed image grid smooth and invertible. The shape-based regularization term works to keep the resulting deformation anatomically reasonable when regions of interest are present in the reference image. Finally, the penalty term added when using controlling structures aims to deform the selected structure in the reference image to the corresponding one in the target image. The ANACONDA algorithm is available in the commercial treatment planning system RayStation (RaySearch Laboratories AB, Stockholm, Sweden). The algorithm can be used for many body sites and it is capable of having a good performance even in cases were image intensity alone is not enough to solve the problem thanks to the usage of the anatomical information of contoured image sets. The image similarity metric that uses ANACONDA is the correlation coefficient and can be applied to an image set of CT and CBCT (Cone Beam Computed Tomography).

Appendix B

Experiments & Results

B.1 Hyperparameter search table

The following table shows the different grid search configurations tested. For some runs, the values of the loss and Dice coefficients do not appear or are very close to 0. This is because for that test the residual connection configuration resulted in the model not learning at some times. Therefore, it was modified in other tests and this modification is stated as TRUE-MOD. Finally, for the loss column, the 5 lowest values are highlighted in green. For the Dice coefficient columns, the 5 models with the highest Dice scores are highlighted in green. It is important to keep in mind that these validation results are obtained on images of the same image resolution as the training images.

lr	Splines interpolation	loss	epoch: optimizer	n_filters	residual	convs_block	val_loss	backgroun	rfemur	bladder	rectum	lfemur	prostate
0,1	1	L2	150 Adagrad	32,64,128,256,512	FALSE	2	1,9668	0,9939	0,8936	0,9391	0,8566	0,8866	0,8811
0,1	1	L2	150 Adagrad	32,64,128,256,512	FALSE	2	0,1114	0,9946	0,8836	0,9466	0,8910	0,8717	0,9141
0,1	1	L2	150 Adagrad	32,64,128,256,512	FALSE	2	0,3590	0,9954	0,9059	0,9533	0,9127	0,9102	0,9037
0,5	1	L2	150 Adagrad	32,64,128,256,512	FALSE	2	0,3720	0,9950	0,8942	0,9501	0,8991	0,9044	0,9029
1	1	L2	150 Adagrad	32,64,128,256,512	FALSE	2	0,7697	0,9941	0,8790	0,9397	0,8700	0,8829	0,8883
0,1	1	L2	150 Adagrad	32,64,128,256,512	TRUE	2	0,2320	0,9955	0,9082	0,9567	0,9064	0,9178	0,9146
0,1	1	MAE	149 Adagrad	32,64,128,256,512	TRUE	2	0,3679	0,9955	0,9150	0,9545	0,9047	0,9110	0,9046
0,1	1	MSE	148 Adagrad	32,64,128,256,512	TRUE	2 stopped due to model not learning							
0,1	1	L2	150 Adagrad	32,64,128,256,512	FALSE	2	0,3460	0,9951	0,8996	0,9486	0,8943	0,9061	0,9085
0,1	1	MAE	150 Adagrad	32,64,128,256,512	FALSE	2	0,3526	0,9957	0,9187	0,9581	0,9175	0,9166	0,9160
0,1	1	MSE	150 Adagrad	32,64,128,256,512	FALSE	2	0,2579	0,9955	0,9080	0,9558	0,9083	0,9134	0,9102
0,1	1	L2	150 Adagrad	32,64,128,256,512	FALSE	4	0,5514	0,9944	0,8810	0,9434	0,8844	0,8834	0,8757
0,1	1	MAE	150 Adagrad	32,64,128,256,512	FALSE	4	0,3908	0,9954	0,9113	0,9531	0,9111	0,9042	0,9156
0,1	1	L2	200 Adadelta	32,64,128,256,512	FALSE	2	0,0867	0,9953	0,8946	0,9483	0,9033	0,8994	0,9157
0,001	1	L2	150 Adam	32,64,128,256,512,1024	FALSE	2	0,1217	0,9948	0,8965	0,9432	0,8917	0,8879	0,9099
0,1	1	MSE	150 Adagrad	32,64,128,256,512	FALSE	4	0,4479	0,9945	0,9033	0,9467	0,8888	0,8930	0,8945
0,1	1	L2	150 Adagrad	32,64,128,256,512	TRUE	4	########	0,9656	0,0000	0,0000	0,0000	0,0000	0,0000
0,1	1	MAE	150 Adagrad	32,64,128,256,512	TRUE	4	0,3466	0,9956	0,9215	0,9549	0,9046	0,9142	0,9004
0,1	1	MSE	150 Adagrad	32,64,128,256,512	TRUE	4	########	0,9640	0,0000	0,0000	0,0000	0,0000	0,0000
0,1	1	L2	150 Adagrad	32,64,128,256,512	TRUE-MOD	2	0,1775	0,9958	0,9222	0,9582	0,9131	0,9177	0,9112
0,1	1	MAE	150 Adagrad	32,64,128,256,512	TRUE-MOD	2	0,3178	0,9958	0,9170	0,9580	0,9142	0,9213	0,9181
0,1	1	MSE	150 Adagrad	32,64,128,256,512	TRUE-MOD	2	0,1954	0,9958	0,9190	0,9580	0,9162	0,9111	0,9168
0,1	1	L2	150 Adagrad	32,64,128,256,512	TRUE-MOD	4	0,3976	0,9946	0,9047	0,9464	0,8698	0,9165	0,8891
0,1	1	MAE	150 Adagrad	32,64,128,256,512	TRUE-MOD	4	0,3374	0,9955	0,9112	0,9574	0,9081	0,9203	0,9201
0,1	1	MSE	150 Adagrad	32,64,128,256,512	TRUE-MOD	4	0,2049	0,9956	0,9126	0,9572	0,9081	0,9058	0,9127
0,1	1	L2	700 Adagrad	32,64,128,256,512	TRUE-MOD	2	0,1259	0,9961	0,9241	0,9607	0,9215	0,9237	0,9218
0,1	1	L2	700 Adagrad	32,64,128,256,512	TRUE-MOD	2	0,0401	0,9960	0,9121	0,9615	0,9229	0,8911	0,9255
0,1	1	L2	700 Adagrad	32,64,128,256,512,1024	TRUE-MOD	2	0,0373	0,9962	0,9062	0,9614	0,9260	0,9059	0,9365
0,1	1	L2	700 Adagrad	32,64,128,256,512,1024	TRUE-MOD	2 stopped due to model not learning							
0,1	1	L2	700 Adagrad	32,64,128,256,512,1024	FALSE	2	0,1373	0,9960	0,9228	0,9604	0,9189	0,9181	0,9190
0,1	1	L2	700 Adagrad	32,64,128,256,512,1024	FALSE	2	0,0630	0,9958	0,9014	0,9595	0,9092	0,9116	0,9266
0,7	1	L2	1500 Adagrad	32,64,128,256,512,1024	FALSE	2	0,0529	0,9958	0,9083	0,9549	0,9277	0,8995	0,9267
0,05	1	L2	1500 Adagrad	32,64,128,256,512,1024	FALSE	2	0,0420	0,9964	0,9217	0,9654	0,9268	0,9077	0,9400
0,3	1	L2	1500 Adagrad	32,64,128,256,512,1024	FALSE	2	0,0428	0,9962	0,9090	0,9631	0,9215	0,9119	0,9327
0,1	1	L2	4458 Adagrad	32,64,128,256,512,1024	FALSE	2	0,0273	0,9959	0,9162	0,9599	0,9142	0,9029	0,9422
0,1	3	L2	700 Adagrad	32,64,128,256,512,1024	FALSE	2	0,0634	0,9965	0,9165	0,9670	0,9304	0,9069	0,9491
0,001	3	L2	700 Adam	32,64,128,256,512,1024	FALSE	2	0,0761	0,9963	0,9111	0,9614	0,9301	0,8940	0,9547
0,001	3	L2	1500 Adam	32,64,128,256,512,1024	FALSE	2	0,0685	0,9962	0,9023	0,9657	0,9198	0,9041	0,9371
0,1/1,0	3	L2	1500 Adagrad	32,64,128,256,512,1024	FALSE	2	0,0448	0,9970	0,9184	0,9729	0,9429	0,9172	0,9524
0,1	3	L2	3000 Adagrad	32,64,128,256,512,1024	FALSE	2	0,0475	0,9966	0,9230	0,9655	0,9273	0,9197	0,9443
0,1/0,01	3	L2	2200 Adagrad	32,64,128,256,512,1024	FALSE	2	0,0351	0,9972	0,9199	0,9727	0,9510	0,9203	0,9691
0,001/0,0001	3	L2	2200 Adam	32,64,128,256,512,1024	FALSE	2	0,0427	0,9971	0,9166	0,9734	0,9487	0,9195	0,9603
0,1	3/1 random	L2	5000 Adagrad	32,64,128,256,512,1024	TRUE-MOD	2	0,0219	0,9973	0,9261	0,9747	0,9506	0,9249	0,9664
0,1	1/3 change at epoch 250	L2	5000 Adagrad	32,64,128,256,512,1024	TRUE-MOD	2	0,0423	0,9969	0,9293	0,9688	0,9376	0,9175	0,9529
0,1/0,3	change 1/3 change at epoch 250	L2	5000 Adagrad	32,64,128,256,512,1024	TRUE-MOD	2	0,0289	0,9967	0,9233	0,9696	0,9335	0,9324	0,9369
0,1	3/1 random + no deformat	L2	5000 Adagrad	32,64,128,256,512,1024	TRUE-MOD	2	0,0232	0,9977	0,9425	0,9761	0,9598	0,9412	0,9553
0,1	3/1 random + no deformat	L2	4000 Adagrad	32,64,128,256,512,1024	TRUE-MOD	2	0,0695	0,9973	0,9441	0,9753	0,9454	0,9445	0,9515
0,1	3/1 random	L2	4000 Adagrad	32,64,128,256,512,1024	TRUE-MOD	2	0,0690	0,9970	0,9363	0,9701	0,9407	0,9430	0,9437
0,1	3	L2	1200 Adagrad	32,64,128,256,512,1024	TRUE-MOD	2	0,4314	0,9945	0,8842	0,9454	0,8869	0,8956	0,8770

Figure B.1: Results of the hyperparameter search.

B.2 CT-CT Displacement Analysis

To better understand the behavior of the neural network and be able to detect in which cases it is failing, further experiments needed to be executed. The analysis of the performance of the 0107 CT-CT registration model on different kind of displacements is presented in this section.

B.2.0.1 Dice-Displacement Analysis

Once obtained the resulting Dice scores it was interesting to see the relationship between the average displacement of the ground truth deformation field and the resulting Dice scores per region of interest. For this analysis, the average displacement is calculated as the average Euclidean distance of the deformation field vectors. In Figure B.2 the Dice-displacement analysis of the CT-CT 0107 model is presented. Our expectation was to see a linear relationship between the increase of displacement and a decrease of the obtained Dice score. Yet, this is not the case. This is because when generating the ground truth deformation field, the deformation vectors are generated randomly as a combination of a finer and coarser grid. In this way, the deformation field affects in a different manner each ROI, deforming some regions with greater deformations than others. Similarly, the same trends can be seen in Figure B.3 where the Dice-displacement analysis for ANACONDA algorithm is presented. Surprisingly, both the deep learning model and ANACONDA work in a similar way against different magnitudes of average displacement. In both figures, each orange dot represents the Dice coefficient of the corresponding ROI of a test sample. In the same way, the blue dots represent the initial Dice score of each test sample and ROI before any registration was performed.

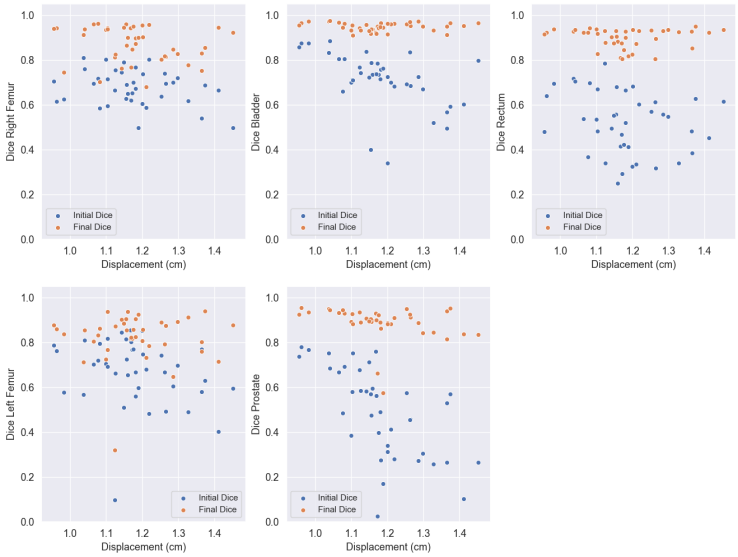


Figure B.2: Ground truth displacement-dice analysis. In this figure the relationship between the initial ground truth displacement and the obtained ground truth score is presented for the CT-CT full resolution model.

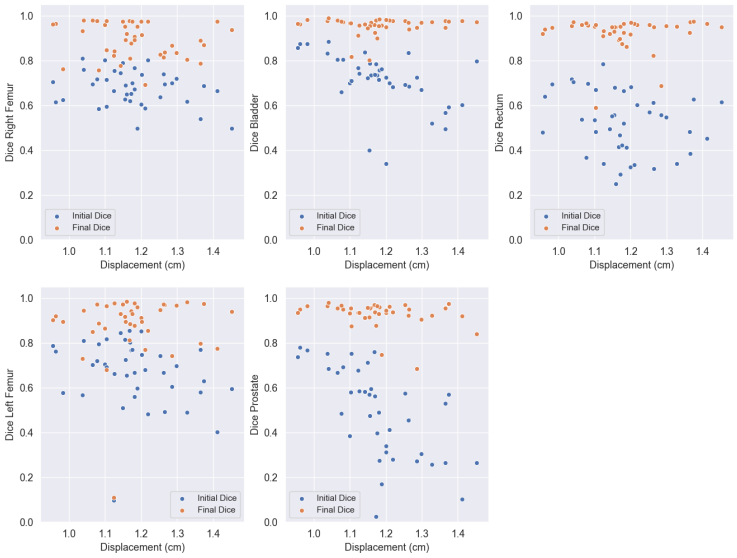


Figure B.3: Ground truth displacement-dice analysis. In this figure the relationship between the initial ground truth displacement and the obtained ground truth score is presented for CT-CT ANACONDA results.

B.2.0.2 Deformation Analysis

In order to test which kind of deformations can affect the model performance, new test data was created by generating deformation vector fields in the same way as presented in Section 2.4, and multiplying it by a factor of 0.25, 0.5, 2, and 4. In this way, fields containing the same distribution of deformations as the ones used in the previous tests are obtained. The goal of this experiment was to test the robustness of the model when facing fields with greater and smaller deformation scales. The model performance was analyzed by comparing the improvement of the Dice coefficient, shown in Figure B.4 where the blue boxes represent the initial Dice scores and the orange ones represent the distribution of Dice scores after applying the predicted deformation field. Additionally, the prediction error is analyzed in Figure B.5 where the blue boxes are the distributions of the ground truth field average displacement, and the orange ones are the distributions of the prediction error. For both plots, factor 1 is the ground truth displacement distribution of the training images. The results show that the model is able to predict the right deformation when facing images containing lower or equal distributions than the training ones. Yet, the model is not able to predict deformations with higher deformation scales than the training ones.

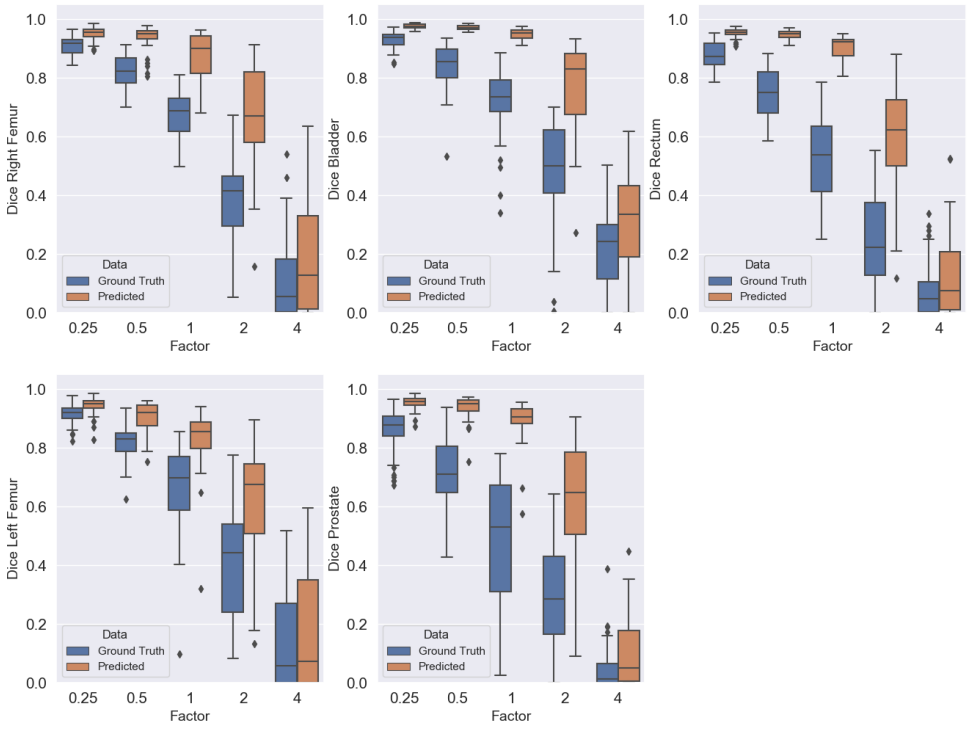


Figure B.4: Dice score analysis from testing the 0107 CT-CT model with deformed images deformed by fields scaled by a factor of 0.25, 0.5, 2, and 4. The initial Dice scores are represented in blue. The orange boxes represent the final Dice scores after applying the model.

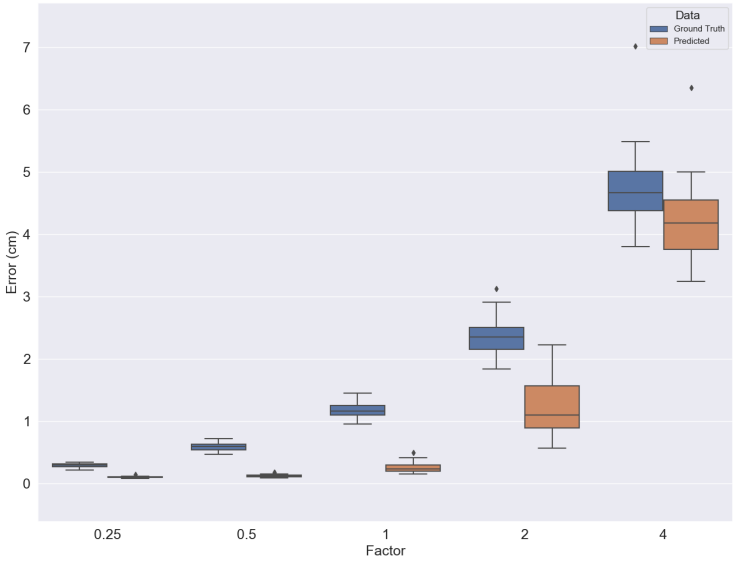


Figure B.5: Displacement-error analysis from testing the 0107 CT-CT model with deformed images with fields scaled by a factor of 0.25, 0.5, 2, and 4. The blue boxes represent the average ground truth displacement, while the orange boxes represent the prediction error per each scale factor .

Given that the model was only trained on deformable deformations, its performance when facing only translations was tested. Test data containing only translations of 0.1, 0.2, 0.5, 1, 1.5, and 3 cm were generated. The resulting Dice scores per translation factor are presented in Figure B.6, where the blue boxes represent the initial Dice scores and the orange ones represent the Dice scores after applying the predicted field. In Figure B.7, the prediction error was analyzed for each translation factor. The initial translation scale is represented in blue, while the prediction error is depicted in orange. The model was able to predict the deformation direction for all tested translation scales but as previously, it could not predict transformations of a higher scale than the ones used for training.

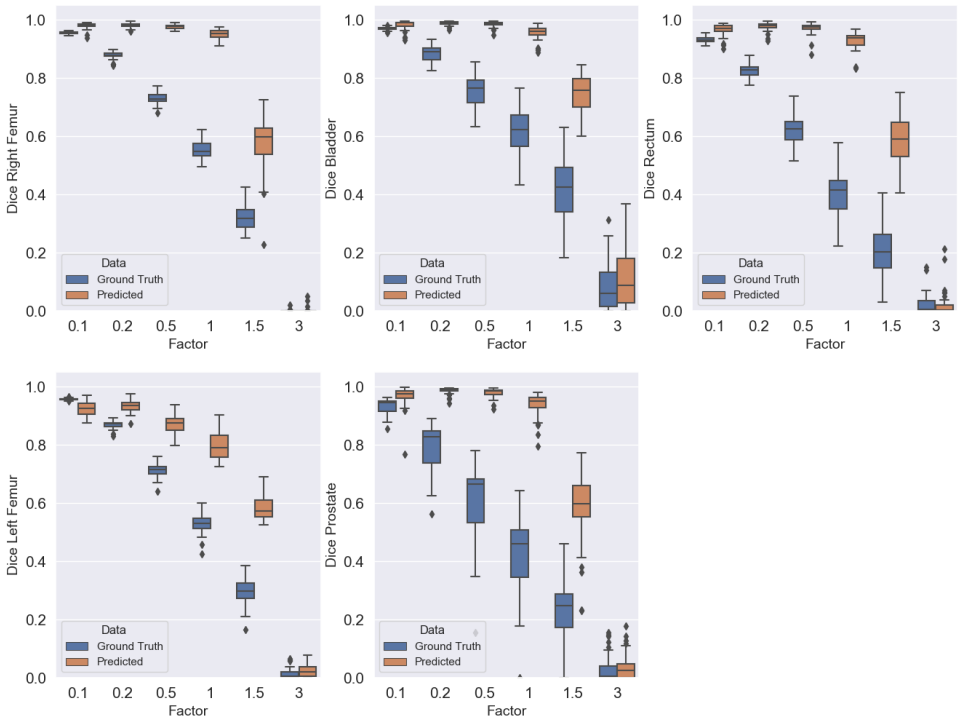


Figure B.6: Comparison of resulting Dice scores from testing the 0107 CT-CT model with images being translated 0.1, 0.2, 0.5, 1, 1.5, and 3 cm. The blue boxes represent the distribution of the initial Dice scores. The orange boxes represent the distribution of the Dice scores obtained after applying the predicted deformation field.

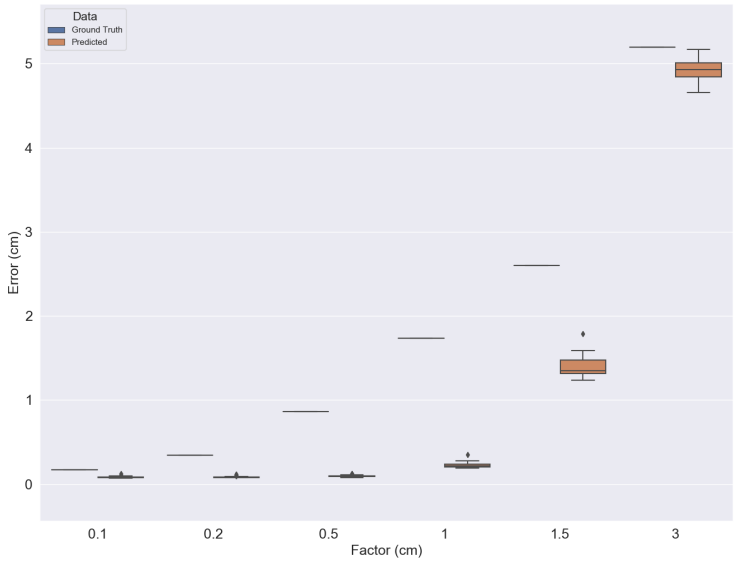


Figure B.7: Dispacement-error analysis from testing testing the 0107 CT-CT model with images being translated 0.1, 0.2, 0.5, 1, 1.5, and 3 cm. The blue boxes represent the distribution of the ground truth deformation fields' average displacement. The orange boxes represent the average prediction error.

TRITA CBH-GRU-2020:097