Universiteit Leiden

# Master Computer Science

An innovative synergy between Evolutionary and Deep Learning Methods for Multi-Objective Deformable Image Registration


| | |
|---|---|
| Name: | Eduard Ruiz Munné |
| Student ID: | 3687988 |
| Date: | August 13, 2024 |
| Specialisation: | Artificial Intelligence |
| 1st supervisor: | Prof. dr. T. Bäck |
| 2nd supervisor: | Dr. T. Alderliesten |


Master's Thesis in Computer Science


Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

# Abstract

Deformable Image Registration (DIR) represents an open challenge in medical imaging, especially when large deformations occur between two images. There are numerous potential applications that could be unlocked given a reliable DIR method, such as the accumulation of delivered dose distributions in radiation therapy treatment planning. Existing methods frequently face a trade-off between achieved accuracy and required runtime. This thesis attempts to address this issue by creating a hybrid method that combines two existing methods: DL-MODIR, a deep learning-based method, and MOREA, an evolutionary algorithm-based method. The proposed method, as its former methods, uses a multi-objective strategy to explore the balance of different objectives, which differ according to the examined patient. By smartly initializing MOREA's population with DL-MODIR solutions for a warm start, the method benefits from both the fast inference times of DL-MODIR and the realism of the biomechanical mesh model of MOREA. The hybrid method is evaluated on two datasets: a synthetic problem involving a shrinking sphere, and abdominal CT scans in which the bladder is full on the source image and empty on the target image. Results indicate that the hybrid method is quantitatively comparable to or significantly better than MOREA exploiting the guidance objective, highlighting the potential of hybrid methods in DIR.

**Keywords:** deformable image registration, deep learning, evolutionary algorithms, radiation therapy planning, DL-MODIR, MOREA

# Acknowledgements

First, I would like to express my deepest gratitude to Prof. Dr. Thomas Bäck for his interest in supervising this project and for providing valuable feedback. I am also thankful to Dr. Tanja Alderliesten and Prof. Dr. Peter Bosman for giving me the opportunity to undertake this project with CWI and the LUMC and supervising the thesis.

Finally, I extend my thanks to Georgios Andreadis for his patience and effort in introducing me to the topic. This project would not have been possible without his invaluable guidance and mentorship.

# List of Acronyms

**BT** Brachytherapy. 1, 2

**CT** Computed Tomography. 2, 3, 37, 39, 64

**DIR** Deformable Image Registration. 2–8, 11–14, 18, 21, 24–26, 63–67

**DL** Deep Learning. 1–3, 7, 11, 12, 16, 20, 21, 24, 65, 67

**DVF** Deformable Vector Field. 7, 10, 12, 21, 24, 26–35, 37, 39, 41, 42, 48, 49, 52, 53, 55, 56, 58, 59, 62–64, 66, 68

**EA** Evoluationary Algorithms. 1–3, 12, 16, 30, 65, 67

**EBRT** External Beam Radiation Therapy. 1, 2

**FEM** Finite Element Model. 7, 10, 13, 17, 24, 27

**FOS** Family Of Subset. 18

**IR** Image Registration. 5, 6

**MRI** Magnetic Resonance Imaging. 64

**OAR** Organ At Risk. 10, 11

**RT** Radiation Therapy. 1, 2, 5, 27, 37, 62, 65, 67

# Contents

# List of Figures

# List of Tables

Chapter 1

# Introduction

Computer vision has impacted medical imaging for decades [20], but with the advent of Deep Learning (DL), the range of opportunities has increased [60]. Their uses include from improved diagnosis to more precise treatments. This thesis leverages DL and Evoluationary Algorithms (EA) methods for enhancing cancer radiation therapy in patients with cervical cancer.

## 1.1 Cervical cancer radiation therapy

Cancer was the second leading cause of death in 2020 with 9.9 million deaths [19] and extensive research is being done for prevention [35] and improvement of current treatments [40]. This thesis focuses on cervical cancer, a disease produced by the abnormal growth of cells in the cervix, the region connecting the uterus with the vagina. It is the third most common cancer in women globally, with 660,000 new cases and 350,000 deaths in 2022 [69], and more than 90% of these cases are caused by infection with human papillomavirus [27] that can be spread by sexual transmission.

According to the European Society of Gynaecological Oncology guidelines, the treatment planning is based on "the comprehensive and precise knowledge of prognostic and predictive factors for the oncological outcome, morbidity, and quality of life" [16]. Hence, each patient receives a different treatment based on their conditions, but a common treatment is Radiation Therapy (RT), consisting of controlled ionizing radiation that induces damage to the DNA of cells. However, since this radiation affects both healthy tissues and tumors, it is necessary to define an acceptable therapeutic ratio, which is defined as the probability of tumor control compared to the probability of unacceptable toxicity. This ratio needs to be administrated with a tolerance of less than 5% deviation [32], requiring precise application of the radiation and keeping track of the delivered radiation dose at each organ. RT can be applied in different forms, such as External Beam Radiation Therapy (EBRT), where radiation is applied externally, or Brachytherapy (BT), which applies radiation internally.

During RT planning the radiation that each organ needs to receive is defined beforehand, with a curative amount for tumoral regions while minimizing the damage to healthy tissues. However, this planning contains uncertainties in dose prescription, tissue tolerances, and the dose distribution delivered to the patient across multiple fractions [63]. This latter is the one that we aim to combat in this thesis. Uncertainty of the delivered dose distribution comes as a result of a change in the

patient's geometry over the course of treatment with deformation of the interior organs. In cervical cancer, this deformation is even more pronounced as the bladder and the rectum can be filled to different degrees, producing a deformation in their surroundings [31]. Additionally, for BT an applicator is introduced as shown in Figure 1.1 to deliver radiation internally, producing that the geometry in BT differs from the ones in planing and EBRT. Thus, the RT plan and delivered dose distribution need to be transformed to the new geometry.

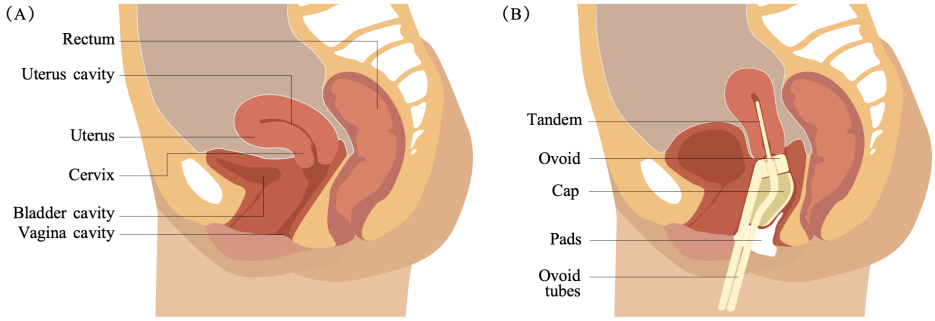



**FIGURE 1.1:** Brachytherapy is performed with an applicator that emits radiation from intracavities. The introduction of the applicator produces a deformation of the uterus cavity and its surroundings. Figure from [52].

Deformable Image Registration (DIR) is the process of modeling the deformation occurring between two images: a source, and a target image. This thesis focuses on EBRT planning, where the treatment is divided into different fractions instead of providing the complete dose in a unique treatment, allowing healthy tissues to recover between treatments [49]. However, during different fractions there are temporal changes and deformations in certain organs, which can potentially reduce the accuracy and robustness of the previous RT plan. Hence, it is necessary to simulate the transformation between the two images to adapt metadata such as the dose administered to each organ to the new geometry, avoiding irradiating healthy tissues and exceeding the therapeutic ratio.

This thesis focuses on DIR in abdominal Computed Tomography (CT) images for its application in RT. Specifically, we predict the deformation occurring between an image with a filled bladder emptied in the second image, which is a non-trivial task since a large deformation occurs during this process, not only deforming the organ but also its surroundings. This is done for planning purposes, as different plans are made to handle different situations: one for an empty bladder, and one for a full bladder. At each RT fraction, practitioners can choose which plan fits best to the current situation. Later, the applied radiation therapy distribution at each fraction can be transformed with DIR for monitoring. Different DIR methods have been developed to solve this problem, dealing with a trade-off between accuracy and registration time. Accurate methods are generally more time-consuming, making them less feasible for clinical practice.

This thesis aims to bridge this gap by using a hybrid method. A previous study developed a hybrid with two DIR methods which shows promising results, outperforming each method independently in terms of accuracy and time [30]. The goal is to further explore this strategy by combining DL and EA methods to bridge the

gap between precision and more realistic results. While DL methods offer fast inference, they may produce less realistic deformations. In contrast, evolutionary methods can simulate body movements more accurately but are more time-consuming. This thesis uses two existing methods in the literature, MOREA and DL-MODIR, an EA method and a DL method respectively. By giving the fast inference results of the DL method to the EA method for a warm start, we aim to have a hybrid method with the precision of evolutionary methods that optimizes faster.

This thesis solves DIR from a multi-objective point of view. A small visual difference between the transformed source image and the target is not enough to define the quality of the deformation. Thus, large and unrealistic deformations can produce results with good visual similarity. Instead, multiple objectives need to be considered for the evaluation. This thesis uses three objectives at the same time: the image intensity difference, the contours or segmentation match, and the deformation magnitude. Consequently, instead of a unique solution, a set of solutions is returned which explores different regions of the objective space where for each given solution does not exist any other solutions that optimize a particular objective without degrading the others. In the end, multiple solutions are given to the clinician, who must select the best solution depending on the patient.

The development of a hybrid method with a multi-objective strategy aims to compute better and more realistic deformations to enhance DIR and have a positive impact on patient outcomes.

## 1.2 Contributions

This thesis presents several contributions in the field of DIR. The core contributions are:

- Develop and implement a hybrid method to bridge the gap of existing DIR methods in the literature.
- Evaluate the effect of a warm start for the MOREA method.
- Evaluate DL-MODIR on a different setting with CT scans and a significant deformation.

## 1.3 Research Questions

We have formulated three research questions that we aim to address to understand how the hybrid method compares to DL-MODIR and MOREA:

- How does the introduced hybrid method quantitatively compare to the MOREA method?
- Does the introduced hybrid method accelerate convergence to high-quality solutions compared with the MOREA method?
- How do the introduced hybrid method solutions qualitatively compare to solutions found by the DL-MODIR method and the MOREA method?

## 1.4 Thesis structure

This thesis is structured as follows. First, in Chapter 2, we introduce DIR and the existing methods in the literature. Chapter 3 describes the developed method and some preliminary experiments. In Chapter 4, the results of the three experiments are reported. Later, in Chapter 5 we discuss the limitations and future work of the implemented hybrid method. Lastly, in Chapter 6 we conclude this thesis.

Chapter 2

# Background

In this chapter, we introduce the problem of Image Registration (IR), and, more specifically, DIR. First, we describe the three main components that comprise a DIR method. Second, we compare several single and multi-objective methods proposed in the literature. Finally, we discuss hybrid methods existing in the literature that combine multiple methods in a pipeline.

## 2.1 Image registration

IR consists of spatially aligning a set of images taken at different time points and with potentially different image modalities (Figure 2.1). Although IR can be applied to sets of more than two images, in this thesis we focus solely on registering a pair of images —a source image $s$ (also known as the moving image) and a target image $t$ (also called the target image). The goal is to find a transformation $\mathbf{u}$ such that $s \circ \mathbf{u} \approx t$, i.e., transforming the source image to match the target image.

IR can be applied to medical images to enable a number of applications, including multi-modality fusion [51] or surgical planning [21]. This thesis focuses on registering medical images taken before and during cervical cancer RT. This is a case where aligning two images separated in time could enable the transfer of spatial metadata, such as delivered radiation dose distribution [14], from one image to the other.

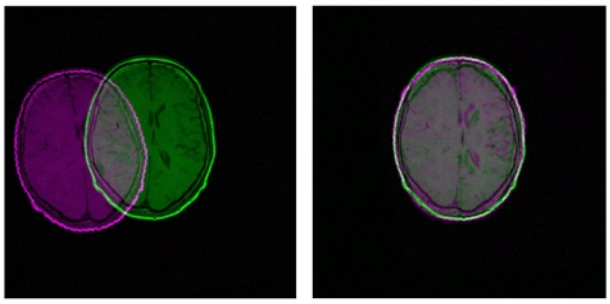


**FIGURE 2.1:** Rigid image registration of two images of a brain, a source MRI scan in magenta and a target CT scan in green. Source: https://www.mathworks.com/help/medical-imaging/ug/medical-image-registration.html

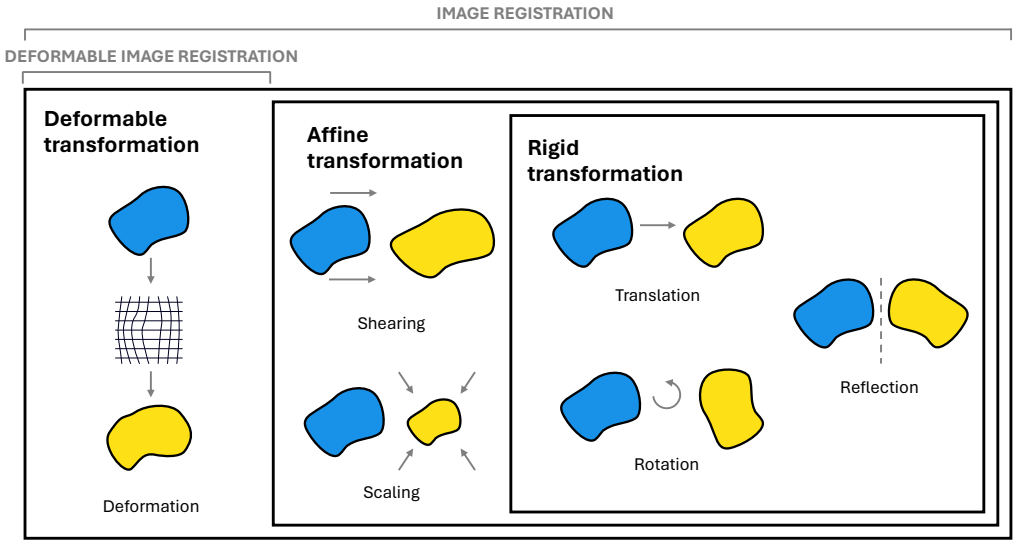



**FIGURE 2.2:** Overview of the different types of transformations that can be found by IR techniques.

As shown in Figure 2.2, simple IR techniques rely on rigid transformations such as rotation, translation, and reflection, to align two images [45, 2]. Affine registration is a more sophisticated method, including scale factors and shears, which provides more degrees of freedom for the alignment. Nevertheless, affine registration is incapable of dealing with local deformations [33].

## 2.2 Deformable image registration

Rigid and affine IR methods are not able to capture local deformations as their transformations are constrained to be linear or affine on the entire image space. DIR methods have been introduced for this purpose, allowing for dense, local, and non-linear transformations. They consist of three main components: its transformation model, registration quality metrics, and optimization method [58]. We discuss several strategies proposed in the literature for each of these components, below.

## 2.3 Transformation models

A transformation model is a representation of a DIR problem which defines the variables that parameterize a transformation. There is a trade-off between computational efficiency and the desired detail of the transformation, so deciding the transformation model to use is a non-trivial decision. More fine-grained transformation models are represented with a larger number of parameters, providing more degrees of freedom but also increasing the model complexity and computational cost. Many different transformation models have been proposed in the literature [58]. In this thesis, however, we focus on three common models, described below.

### 2.3.1 Finite element model

A Finite Element Model (FEM) is a deformable mesh representing a transformation [72] (Figure 2.3a). While the mesh topology remains fixed, the vertices of the mesh (also called nodes) can be moved, subject to folding constraints. Within this model, biomechanical properties such as the elasticity of the modelled organs can be considered. These relate to the physical characteristics of the imaged anatomical structure, which can improve the realism of the model.

Implicitly, this model has a fixed mesh on the target image while the source image mesh is deformed. Alternatively, both meshes, source and target, can be moved at the same time, in a model known as the dual-dynamic [61]. This model makes it possible to flexibly tackle large deformations and cope with the appearance and disappearance of structures between one image and the other. MOREA [4], for instance, uses a dual-dynamic model to capture large deformations with a FEM-based transformation model. This enables two types of registration: unidirectional and symmetric. In unidirectional registration, only one transformation is considered, such that $s \circ \mathbf{u} \approx t$. Alternatively, symmetric registration also considers a second transformation $\mathbf{u}'$ such that $t \circ \mathbf{u}' \approx s$. This can make solutions more physically viable, and resistant to image input order changes [58].

### 2.3.2 B-splines

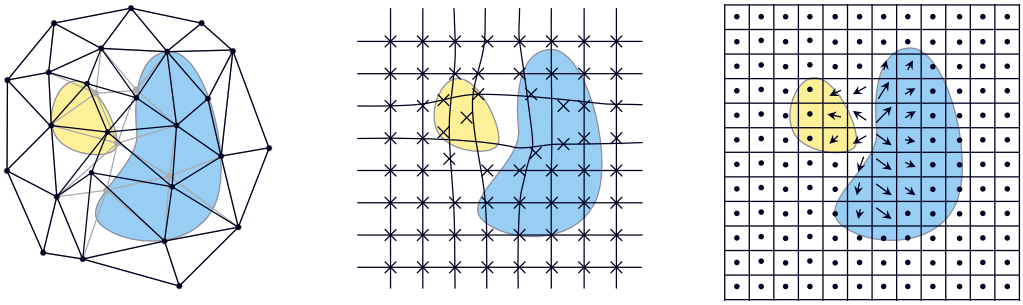
The B-splines transformation model consists of a grid of control points that parameterize transformations defined by B-splines. A representation of B-splines is shown in Figure 2.3b, where crosses represent the control points. B-splines are based on Bézier curves, which are parametric curves modeled by a set of control points. As a result, smoother deformations are generally obtained compared to mesh FEM-based models due to the attenuation that the B-spline function applies to the control point movements. Yet, it has been observed that this model suffers when dealing with large and discontinuous deformations [4].

### 2.3.3 End-to-end

With the advent of DL, a transformation model has emerged which does not require a higher order parametrization of the problem. Instead, a displacement vector is encoded and optimized for each voxel coordinate (Figure 2.3c) generating a Deformable Vector Field (DVF). This DVF contains a mapping of voxel coordinates from one image to the other image. Specifically, an inverse DVF points from target voxel coordinates to source voxel coordinates. Hence, it can represent more fine-grained deformations, but at the cost of an increased search space where many more parameters need to be optimized. Due to the lack of an underlying explicit physical model, it may be more difficult to derive realistic deformations with this transformation model without an additional higher-level representation.

### 2.3.4 Other models

Further methods exist in the literature to represent transformation models. Demons [62], one of the first successful DIR methods, is based on viscous-fluid equations employing diffusion techniques from thermodynamics. Others are based on flows



**(A)** FEM-based model. Two meshes, the original (gray) and the transformed (black).

**(B)** B-spline model representation where each cross represents a control point.

**(C)** DVF in end-to-end models. Points represents the absence of transformation.

**FIGURE 2.3:** Three transformation models proposed in the literature to parametrize the DIR problem: FEM-based, B-splines, and end-to-end.

of diffeomorphisms, invertible DIR problem definitions where both, the function and its inverse, are differentiable, resulting in smooth transformations [28]. A more extensive classification of transformation models available in the literature is offered by Sotiras et al. [58].

## 2.4 Registration quality metrics

Once a transformation model has been defined, it is necessary to be able to judge the quality of a registration defined in this model. Registration quality metrics are indicators that measure the quality of a registration. Multiple metrics exist in the literature to define this quality from different angles.

We will use the following notation: $\Omega$ represents the set of voxels in each image, where each voxel $p \in \Omega$ is a three-dimensional point $p \in \mathbb{N}^3$ representing its spatial position, and $|\Omega|$ is the total count of voxels in the image. For a more compact expression, we define the transformed source image $s' := s \circ \mathbf{u}$, where $\mathbf{u}$ defines the measured transformation. Therefore, the transformed source image at each point $p$ corresponds to the transformation $\mathbf{u}(p)$ applied to the source image, such that $s' := s \circ \mathbf{u}(p)$.

### 2.4.1 Intensity-based metrics

Intensity-based similarity metrics focus on the difference in voxel intensity between the target image $t$ and the transformed source image $s$. These range from very simple metrics, such as the mean squared difference, to more complex, such as the normalized cross-correlation. While intensity metrics are useful to define how similar is the transformed source image compared to the target image, they can fall short in low contrast regions [71, 34].

#### Mean squared difference

The mean squared difference metric averages the quadratic difference in voxel intensity between the target image and transformed source image. Thus, large differences are highly penalized by computing the square of the difference. Finally, the resulting value is averaged for all pixels $\Omega$, ensuring that the metric is normalized for different image sizes.

$$MSD = \frac{1}{|\Omega|} \sum_{p \in \Omega} (s'(p) - t(p))^2 \tag{2.1}$$

#### Normalized cross-correlation

The normalized cross-correlation metric computes the similarity of the intensity values of two images. This makes it possible to assess how well two uncalibrated images correspond [70] since it is capable of handling linear relationships between intensity values in multi-modality applications, therefore being more robust than the mean squared difference [6]. It is mathematically defined as follows:

$$NCC = \frac{\sum_{p \in \Omega} t(p) \cdot s'(p)}{\sqrt{\sum_{p \in \Omega} t(p)^2 \cdot \sum_{p \in \Omega} s'(p)^2}} \tag{2.2}$$

Normalized cross-correlation values range in the interval $[-1, 1]$. Intuitively, a value of 1 represents a perfect match between the images, $-1$ is an inverse match, and no correlation is expressed as 0. There is an extension of this metric that normalizes the score by subtracting the average of each image, called the zero-normalized cross-correlation:

$$ZNCC = \frac{\sum_{p \in \Omega} (t(p) - \bar{t}) \cdot (s'(p) - \overline{s'})}{\sqrt{\sum_{p \in \Omega} \left(t(p) - \bar{t}\right)^2 \cdot \sum_{p \in \Omega} \left(s'(p) - \overline{s'}\right)^2}} \tag{2.3}$$

where $\bar{t} = \frac{1}{|\Omega|} \sum_{p \in \Omega} t(p)$ and $\overline{s'} = \frac{1}{|\Omega|} \sum_{p \in \Omega} s'(p)$ are the average of $t$ and $s'$ respectively. As a result, this metric is even more robust, addressing intensity shifts by subtracting the mean of each image.

### 2.4.2 Energy-based metrics

Relying solely on intensity values can result in seemingly equal but unrealistic transformations caused by excessive local deformations that are not directly visible in the transformed source image. It is therefore important to add a regularization term that minimizes the exerted deformation. For this purpose, energy-based metrics are used to assess the plausibility of the transformation. In this thesis, we refer to the magnitude objective the one that uses energy-based metrics.

#### Spatial gradient displacement

To measure the smoothness in the transformation, the spatial gradient displacement computes displacement changes of neighboring voxels [7] as follows:

$$\sum_{p\in\Omega} \left\| \left( \frac{\partial \mathbf{u}(p)}{\partial x}, \frac{\partial \mathbf{u}(p)}{\partial y}, \frac{\partial \mathbf{u}(p)}{\partial z} \right) \right\|^2 \tag{2.4}$$

where $\mathbf{u}(p)$ contains the displacement vector at point $p$, and the partial derivative is approximated as follows:

$$\frac{\partial \mathbf{u}(p)}{\partial x} \approx \mathbf{u}(p_x + 1, p_y, p_z) - \mathbf{u}(p_x, p_y, p_z) \tag{2.5}$$

High values are associated with locally heterogeneous deformations, which can be unrealistic and therefore of lower quality.

#### Bending energy

The bending energy [53] is another energy-based metric that ensures smoothness in the transformation [59]. It is computed as the second-order derivative of the DVF ($\mathbf{u}$), estimated as follows:

$$\frac{1}{|\Omega|} \sum_{p\in\Omega} \left[ \left( \frac{\partial^2 \mathbf{u}}{\partial x^2} \right)^2 + \left( \frac{\partial^2 \mathbf{u}}{\partial y^2} \right)^2 + \left( \frac{\partial^2 \mathbf{u}}{\partial z^2} \right)^2 + 2\left( \frac{\partial^2 \mathbf{u}}{\partial xy} \right)^2 + 2\left( \frac{\partial^2 \mathbf{u}}{\partial xz} \right)^2 + 2\left( \frac{\partial^2 \mathbf{u}}{\partial yz} \right)^2 \right] \tag{2.6}$$

The second derivative measures the amount of bending and curvatures that the DVF contains. Therefore, this term is zero for affine transformations, penalizing only non-affine transformations [66].

#### Hooke's law energy

In FEM-based transformation models, the deformation magnitude is equivalent to the amount of energy required to perform a transformation from the source mesh to the target mesh. Hooke's law states that the force needed to compress or expand a spring is equivalent to the product of the distance of displacement by a constant characteristic from the spring. In a FEM-based model, edges are modelled as springs and the constant value corresponds to the specific elasticity constant of the tetrahedron (defined by the organ it models). The energy metric is computed by considering all edges $e_s$ and $e_t$ belonging to the source and target image respectively for each tetrahedron $\delta \in \nabla$. As defined in the dual-dynamic transformation model of MOREA [4], a total of 10 edges per tetrahedron are used [3], together with the tetrahedron elasticity constant $c_\delta$. The energy is computed as follows:

$$\frac{1}{10|\nabla|} \sum_{\delta\in\nabla} \left[ \sum_{(e_s,e_t)\in E_\delta} c_\delta (||e_s|| - ||e_t||)^2 \right] \tag{2.7}$$

### 2.4.3 Segmentation-based metrics

Another approach consists of segmenting Organ At Risk (OAR). Segmentation-based metrics measure the overlapping region between the transformed source and the target image occupied by an organ. Many organ segmentation methods

are studied in the literature, ranging from automatic algorithms utilizing computer vision algorithms [24], to manual segmentation, performed by experts. While manual segmentation is time-consuming and requires expertise, it often yields more accurate results [67], being considered as the gold standard. In contrast, solely relying on automatic segmentation can introduce errors that may be accumulated in the registration process, specially in the case where boundary information is not visually distinguishable [67]. Thus, hybrid strategies, combining automated and manual segmentation in a pipeline [48, 8], can lead to better accuracy.

#### Dice similarity

The dice similarity coefficient computes the proportion of shared regions between two segmentations considering the total area that they occupy together.

$$DSC = \frac{2|t \cap s'|}{|t| + |s'|} \tag{2.8}$$

Higher dice similarity coefficient values are related to larger overlap between both segmentations, hence indicating a better registration for that specific OAR.

#### Contour metrics

Segmentations can also be used to generate contours, defined by a set of points that outline the segmented OAR. In a DIR problem, we consider two sets of points $C_s$ and $C_t$ for both the source and target image, respectively. To measure the match between the transformed source image set and the target image set, the distance between them is measured and minimized. MOREA [4], for instance, uses the Chamfer distance as a metric. This is the longest minimum distance between a point in the transformed source image and any point in the target image.

Nonetheless, the presence of spatially distributed outliers or a larger number of points in one of the sets, due to different-sized OAR in each image, are challenges that contour-based metrics face [15].

### 2.4.4 Landmark-based metrics

Landmark points are anatomically relevant points corresponding to anatomical structures defined in both images (source and target), intended to help the system identify identical structures across both volumes. Their annotation can require less human effort than OAR segmentation and serve as a guide for transformation.

Recently, automatic methods for landmark generation based on DL approaches have been developed, offering a higher number of points albeit with less accuracy [23, 26]. As concluded by Grewal et al., a larger number of less accurate landmarks may prove more helpful than a lower amount in highly deformed regions, making automatic methods a promising strategy to acquire landmarks.

Landmark metrics describe the distance between these labeled points between the target and the transformed source image. The final quality value is the average point-pair distance:

$$\frac{1}{|\Omega|} \sum_{p \in \Omega} d(s'(p), t(p)) \tag{2.9}$$

A distance metric is necessary to qualify the difference between two points, with the Euclidean distance being the most used.

## 2.5 Optimization methods

The definition of a transformation model together with registration quality metrics constitute the definition of a DIR problem. The next step is to optimize the parameters of the transformation model to improve registration quality metrics. Many optimization methods are proposed in the literature, but in this thesis we will focus on two of them: deep learning and evolutionary algorithms.

### 2.5.1 Deep learning

DL is a continuous optimization method in which variables are in the form of real values. In this case, instead of adjusting the transformation model parameters for each pair of images (source and target) as other methods do, a common representation for all pairs in a training dataset is learned. This allows replacing costly optimization methods, performed individually for each sample, with a unique and global function that encompasses all instances in a shared distribution, therefore accelerating the process. After a function is learned in the training phase, new images are registered by evaluating the learned function. However, it is necessary a large dataset and that new instances fall in the same data distribution as training images.

There are two main DL approaches depending on the learning process:

- **Supervised optimization:** this approach requires labeled data. Apart from the source and target images, a DVF is also needed. As the true DVF rarely is available, artificial DVFs can be used for training [57].
- **Unsupervised optimization:** this approach does not require a DVF. Instead, only the pair of images is needed by the model. To distinguish good from poor deformations, the model measures one or various registration quality metrics to evaluate the transformation. As a result, parameters are updated in the direction where those metrics are optimized. VoxelMorph [7] is an example of an unsupervised DL optimization method, which is also the basis for the DL model used in this thesis.

### 2.5.2 Evolutionary algorithms

EA are heuristic methods based on the theory of evolution and natural selection for optimization. They use mechanisms inspired by biological evolution, such as mutation and recombination, to perform changes to a population of solutions. Only the best solutions, evaluated based on a fitness function (equivalent to the loss function in DL), are recombined in future generations. This method is found to converge slowly but can avoid falling into local optima [13]. MOREA [4], one of the methods used in this thesis, is based on EA to optimize a DIR problem.

## 2.6 Single-objective deformable image registration

Most methods in the literature focus on single-objective DIR using a single optimization objective for registration. Earlier DIR methods mainly focus on intensity-based registrations aiming to minimize voxel intensity differences between the target and the transformed source image. In 1996, Demons emerges as a novel method employing diffusion techniques from thermodynamics. Nevertheless, Demons, like most of its successors, is constrained to a single objective. Two alternatives are evaluated in the original paper [62], one for intensity metrics and another for contour metrics, both of them evaluated independently. Some advancements are later introduced by ANTs SyN, leveraging cross-correlation metrics between images for intensity matching using diffeomorphic transformation models [6]. These mark progress in image-based registration by exploiting the guarantee of invertibility of diffeomorphic transformation models.

However, relying solely on a single optimization objective is prone to overfitting. Consequently, multiple registration quality metrics can be grouped in a single optimization objective giving a weight to each metric according to its importance. For instance, in the case of dealing with three registration quality metrics: intensity, energy, and landmarks, three different values $w_1$, $w_2$, and $w_3$ need to be defined *a priori*. As a result, the optimization objective would be represented using linear scalarization as follows:

$$\mathcal{L} = w_1 \text{Intensity} + w_2 \text{Energy} + w_3 \text{Landmark} \tag{2.10}$$

where $w_i \in \mathbb{R}$ are real-valued weights, with $i \in 1, ..., k$, and $k$ being the number of registration quality metrics.

Methods such as ANACONDA [68] combine both intensity and contour-based metrics, in a single objective using linear scalarization with weights defined *a priori*. By considering more than one metric, these methods aim to achieve more accurate registrations than relying on a single criterion.

The Elastix toolbox [36] utilizes B-spline transformation models combining more than one registration quality metric using linear scalarization. However, it suffers when dealing with large deformations compared to other methods [4].

Other methods introduced energy metrics, but this metric itself is insufficient to guarantee realistic deformations. To address this challenge, energy functions with biomechanical properties were introduced. These use energy-based metrics as a regularization parameter to obtain more physically plausible solutions while maintaining a good match between the target and transformed source image. Incorporating this regularization step yields more authentic deformations compared to relying solely on intensity or contour metrics [50]. Different methods have been developed which use a FEM-based model and the elasticity of the tissues. This elasticity can either be homogeneous, ensuring uniform deformation across all organs [1], or specific, assigning different elasticities to individual organs [11].

MORFEUS [11] combines these biomechanical properties with contour-based metrics to achieve better and more realistic deformations. This method integrates three different commercial software: a FEM pre-processor, a finite element analysis system, and a treatment planning system.

Nevertheless, in general, single-objective optimization using linear scalarization is complex and time-consuming. Finding an optimal combination of objective weights is challenging. Manual and grid search methods have been evaluated, but are slow and inaccurate. To address this challenge, more efficient methods have been studied. For instance, iterative optimization methods [47] have been shown to perform better than regular grid search.

## 2.7 Multi-objective deformable image registration

It has been shown that the optimal set of optimization objective weights may vary depending on the pair of images being registered [47]. Different registrations can be superior in one optimization objective but not in the others. In this case, we can use multi-objective optimization to find a set of solutions that is Pareto-optimal with respect to the objectives. In this section, we define the concept of Pareto dominance and describe how it can be applied in the context of DIR.

### 2.7.1 Pareto front

In multi-objective optimization, more than one objective is simultaneously optimized. Sometimes, optimizing one objective has a negative effect on another. This is where multi-objective optimization takes place. Instead of having a unique solution, a Pareto front or set of non-dominated solutions is obtained [43].

Considering $k$ objectives ($k \geq 2$ in multi-objective optimization) that need to be minimized $\min_{x \in X} f_1(x), f_2(x), ..., f_k(x)$, a solution is called to be non-dominated if none of the objectives can be improved without degrading at least one of the other objectives. In the case of two solutions $x_1, x_2 \in X$; $x_1$ is said to Pareto dominate $x_2$ —mathematically noted as $x_1 \succeq x_2$ —if the following two conditions are met:

$$\begin{array}{ll} \forall i \in \{1, ..., k\} & f_i(x_1) \leq f_i(x_2), \quad \text{and} \\ \exists i \in \{1, ..., k\} & f_i(x_1) < f_i(x_2) \end{array} \tag{2.11}$$

Therefore, the Pareto front is defined as the set of non-dominated solutions (Figure 2.4) and a solution is called Pareto optimal if no solution dominates it.

Various metrics have been proposed to evaluate and compare different Pareto fronts. In this thesis we will use the hypervolume, a metric extensively used in multi-objective optimization to indicate the quality of a Pareto front.

### 2.7.2 Hypervolume

The hypervolume of a Pareto front is the area –in a two-dimensional objective space– that the approximated Pareto front covers with respect to a user-defined reference point $r$ that sets the limits (Figure 2.5) [12]. In a three-dimensional objective space, the hypervolume is a volume.

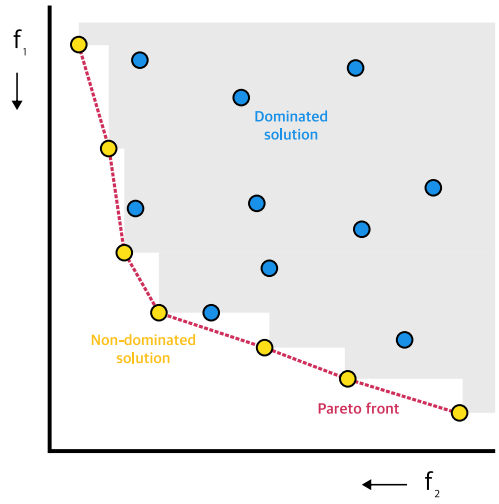



**FIGURE 2.4:** Pareto front representation in a 2D objective space, where $f_1$ and $f_2$ are two objectives that need to be minimized. Colored in gray is the region dominated by each solution (top-right of the solution). Non-dominated solutions (yellow) form the Pareto front that dominates the dominated solutions (blue).

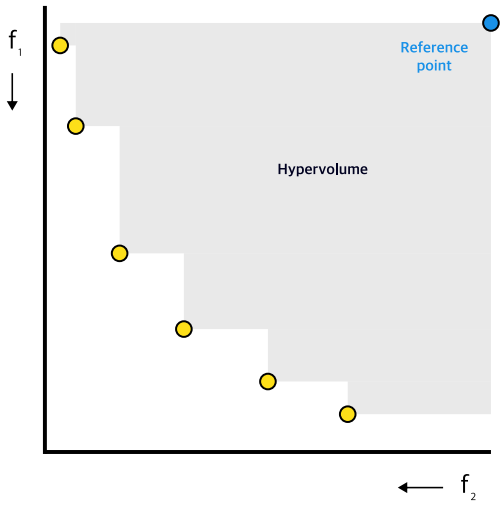
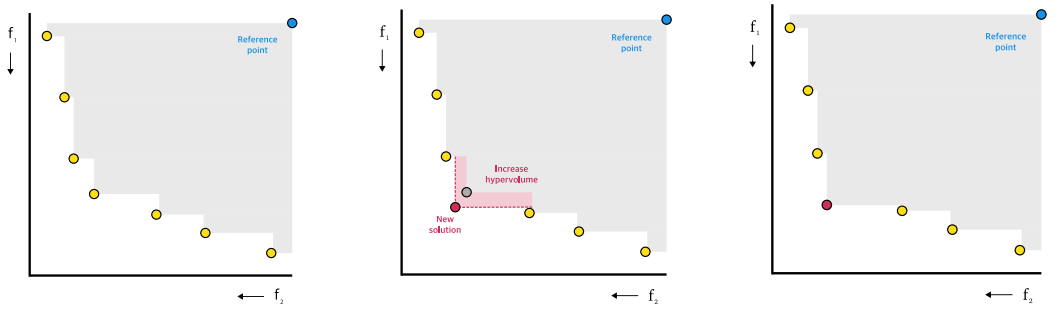



**FIGURE 2.5:** Hypervolume of an approximated Pareto front with respect to two objectives $f_1$ and $f_2$. A user-defined reference point sets the limits of the area that the approximated Pareto front occupies.

When comparing two Pareto fronts, it is said that a Pareto front dominates another when the former has a higher hypervolume than the latter. This can occur for two reasons:

- A new solution dominates a previous non-dominated solution. In this case, the hypervolume increases since the new solution encompasses the previous non-dominated solution inside its hypervolume (Figure 2.6).

- The Pareto front is more diverse and solutions are more spread in the objective space. Thus, a higher area is occupied by the diverse solutions (Figure 2.7).

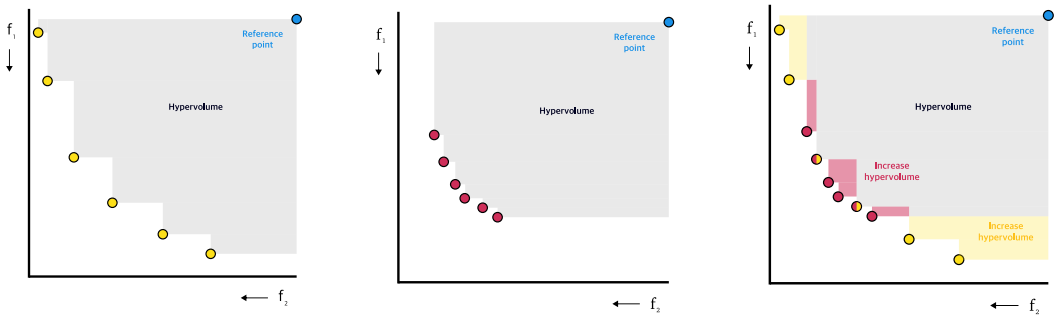



**(A)** Original approximated Pareto front.

**(B)** New non-dominated solution increases the hypervolume.

**(C)** New approximated Pareto front after adding the new non-dominated solution.

**FIGURE 2.6:** The hypervolume of an approximated Pareto front increases after adding a new non-dominated solution that dominates a non-dominated solution from the original Pareto front.




**(A)** Spread Pareto front over the objective space.

**(B)** Compressed Pareto front over the objective space.

**(C)** Hypervolume comparison between the spread and the compressed Pareto front.

**FIGURE 2.7:** Spread Pareto fronts have a higher hypervolume, as a larger area of the objective space is occupied.

### 2.7.3 Multi-objective methods

Multi-objective methods generate a set of non-dominated solutions from which users need to select their desired solution for each specific case *a posteriori* [46].

Recently, various methods with Multi-Objective Deformable Image Registration (MODIR) have emerged to tackle this challenge, employing different optimization methods. These include EA with mesh grids [4] or B-splines [5], and a method using multi-objective DL [25].

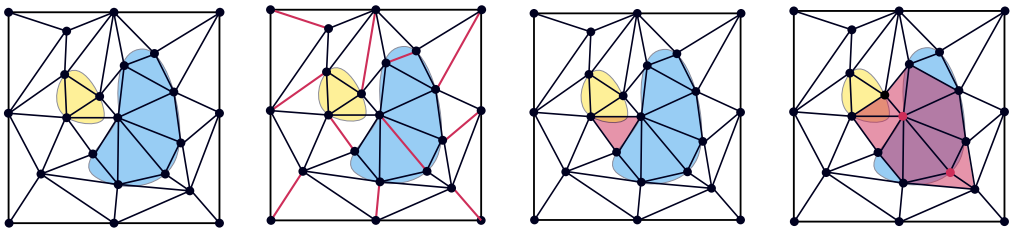
This thesis is based on two of them: an evolutionary-based method, MOREA [4] (Section 2.7.4), and a DL method, DL-MODIR (Section 2.7.5). Both are explained in more detail in the following sections and, later on, compared (Section 2.8).

### 2.7.4 MOREA

MOREA is a FEM-based transformation model optimized with an EA that outperforms state-of-the-art methods, such as Elastix and ANTs SyN, in 3 out of 4 evaluated patients when registering large deformations [4].

As a multi-objective method, it uses three different registration quality metrics: the mean squared difference for the intensity, the Chamfer distance for the contours (also referred to as guidance), and Hooke's law for the magnitude. MOREA uses the Multi-Objective Real-Valued Gene-pool Optimal Mixing Evolutionary Algorithm (MO-RV-GOMEA) [10], a multi-objective EA with model-based linkage models, resulting more scalable than other popular methods such as NSGA-II [17] and MAMaLGaM [9], specially when partial deformations, typical of Gray-Box Optimization, are possible.

MOREA subdivides the problem into smaller parts, each of which is a tetrahedron, that together form a tetrahedral mesh. Each small part is a sub-function of the problem (Figure 2.8c) used for partial evaluation, where only modified tetrahedra need to be re-evaluated (Figure 2.8d). Additionally, sub-functions allow to define constraints, e.g. to avoid folds in the mesh. This occurs when one tetrahedra's volume changes its sign. To avoid this happening, larger folds yield to bigger constraints.



**(A)** Mesh grid transformation model. **(B)** Greedy set cover for FOS linkage set generation. **(C)** Definition of a sub-function (red region) for dra partial evaluation. **(D)** Evaluated tetrahe after a FOS (red nodes) is updated.

**FIGURE 2.8:** MOREA mesh representation, FOS linkage set generation and partial evaluations with a tetrahedron being a sub-function. Updating one FOS linkage set causes the neighboring tetrahedra to be re-evaluated.

The tetrahedral mesh shapes the FEM-based transformation model, where each node defines a vertex of a tetrahedron represented in three dimensions: depth, height, and width coordinates $(x, y, z)$. Firstly, the nodes are randomly distributed across the image, reserving a percentage of them to be placed on the contours of the organs to capture their shape. Later, a surface is generated using a Marching Cubes algorithm [41], and the tetrahedral mesh is rendered given the surface mesh using TetGen [55]. Considering that the mesh of an image consists of $N_n$ nodes, it can be represented as an array:

$$\{x_1, y_1, z_1, x_2, y_2, z_2, ..., x_{N_n}, y_{N_n}, z_{N_n}\} \tag{2.12}$$

with a length of $N_n \times 3$ where $x_i, y_i, z_i \in \mathbb{R}$ are the coordinates for each node $i \in \{1, ..., N_n\}$. However, since MOREA uses a dual-dynamic transformation model, two meshes are needed, one for each image. Thus, the generated mesh is duplicated

for the second image and later disrupted. The DIR solution can be expressed with the following array:

$$\{x_1^s, y_1^s, z_1^s, ..., x_{N_n}^s, y_{N_n}^s, z_{N_n}^s, x_1^t, y_1^t, z_1^t, ..., x_{N_n}^t, y_{N_n}^t, z_{N_n}^t\} \tag{2.13}$$

with $x_i^s, y_i^s, z_i^s$ with $i \in \{1, ..., N_n\}$ being the node coordinates for the source image and $x_i^t, y_i^t, z_i^t$ with $i \in \{1, ..., N_n\}$ the node coordinates for the target image. Hence, a DIR solution contains $2 \times N_n \times 3$ variables, one for each node coordinate in each image. Some additional constraints are needed for each node coordinate to ensure that they stay in the image space:

$$x_i = \begin{cases} 0, & x_i < 0; \\ x_i, & 0 \leq x_i \leq \text{depth}; \\ \text{depth}, & x_i > \text{depth}; \end{cases} \tag{2.14}$$

$$y_i = \begin{cases} 0, & y_i < 0; \\ y_i, & 0 \leq y_i \leq \text{height}; \\ \text{height}, & y_i > \text{height}; \end{cases} \tag{2.15}$$

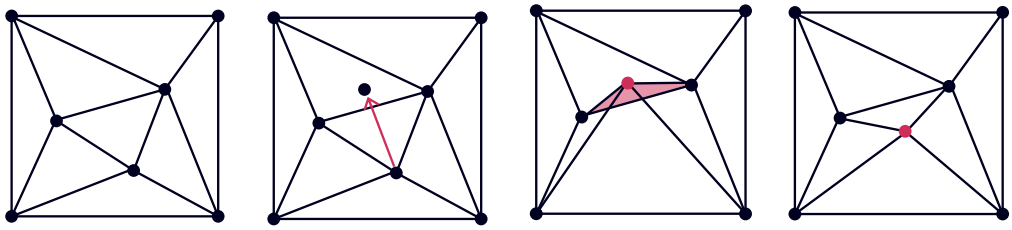
$$z_i = \begin{cases} 0, & z_i < 0; \\ z_i, & 0 \leq z_i \leq \text{width}; \\ \text{width}, & z_i > \text{width}; \end{cases} \tag{2.16}$$

The first step is the population initialization. MOREA is developed to guarantee diversity in the initial population while ensuring no mesh folds that lead to infeasible solutions. A stochastic algorithm generates solutions without folds by using attractors, radial basis functions that attract nodes to move stochastically towards it, generating various solutions that come from the original generated mesh. This method is executed through a multi-step process with incremental adjustments at each stage, therefore avoiding mesh folds while guaranteeing diversity.

After initializing the population, the optimization process takes place. Each solution from the population is mutated and recombined to generate new solutions. Instead of completely modifying it, partial variations are done. This also allows partial evaluations by re-evaluating only the affected sub-functions. MOREA generates sets of dependent variables called Family Of Subset (FOS) that contain dependence relations, which are being jointly updated. For this problem, MOREA takes an edge of the tetrahedron, connecting two nodes as a FOS. A greedy set cover algorithm generates the FOS linkage set (Figure 2.8b), where every node is part of a FOS.

However, when generating new solutions, folds can occur. A repair method is needed to avoid discarding them, especially when they are promising (Figure 2.9). Each node causing a fold is modified by sampling from a Gaussian distribution around it. A total of 64 different solutions are sampled and re-evaluated for each infeasible node. It is selected if a repaired solution is feasible and better than old solutions. Otherwise, it is discarded. This process works similarly to natural selection, where the best solutions are kept and others discarded. As it is known that

elitism helps to converge faster [37], MO-RV-GOMEA monitors an archive of non-dominated solutions during the optimization process. Archived solutions are clustered, with $q$ equal-sized clusters, to divide the entire approximated Pareto front in regions, updating all solutions belonging to the same cluster into a direction of the objective space (see Figure 2.10).



**(A)** Initial solution without folds. **(B)** Solution update causes folds. **(C)** Negative area of one tetrahedron. **(D)** Fold repair method.

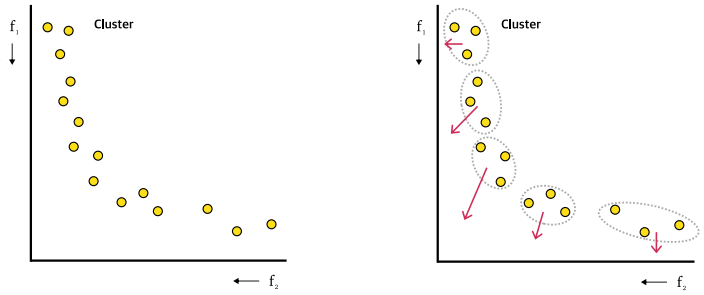
**FIGURE 2.9:** Representation of a mesh fold and how the repair method works in MOREA. Folds cause a negative area of one tetrahedron, requiring a repair method that modifies the node to get a feasible solution.

In this way, at each generation, each cluster optimizes its solution in one direction of the objective space, as shown in Figure 2.10, trying to maximize the hypervolume of the approximated Pareto front. Some additional mechanisms are used for faster convergence. The Anticipated Mean Shift (AMS) allows each cluster to move toward improvement. This direction is defined as the difference between the cluster mean of the current generation and the previous generation (Figure 2.11). Thus, momentum is exerted to ensure an increase in the hypervolume.

This momentum rate is regulated by the distribution multipliers, scaling the variance of the Gaussian distribution in the mutation step. It may happen that when going in the direction of the optimal point (or hypervolume maximization in multi-objective problems), solutions start to collapse (Figure 2.12). Thus, many steps are required to reach the optimal point. In this case, we aim to increase the variance of the Gaussian distribution to converge faster to the optimum. On the other hand, if solutions are around the optimal point, the model is prone to overshoot. Given this case, it is necessary to decrease the distribution multiplier, reducing the variance of the Gaussian to collapse into the optimal point. MO-RV-GOMEA considers both cases, tuning the distribution multiplier for faster convergence.

In some cases, no improvement may occur for several consecutive generations. In that case, a forced improvement moves solutions toward the current best solution. A detailed description of this method can be found in the MOREA paper [4], as well as in the MO-RV-GOMEA paper [10], containing a description of the algorithm MOREA uses for optimization.

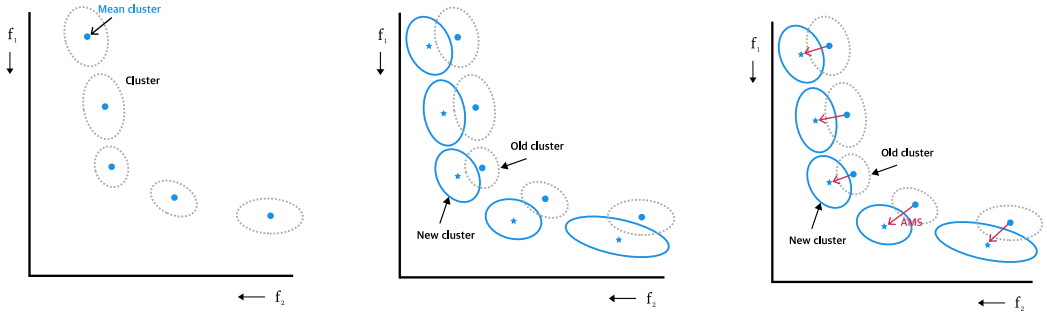
It is found that the energy metric restricts the deformation of solutions. This is a challenge, specially when dealing with large deformations. So, despite the diversity in the solutions is the main purpose of multi-objective approaches, in large deformation problems there is no interest in solutions with low deformation energy. To avoid considering them, an adaptive steering strategy is used every 100 generations to prevent the model from collapsing in small deformations. The front




**(A)** Population of solutions from RV-MO-GOMEA.

**(B)** Solutions clustering, each cluster moves to a different direction.

**FIGURE 2.10:** Clustering process, each cluster moves in a different direction from the objective space.




**(A)** Previous generation clusters.

**(B)** New generation clusters.

**(C)** AMS computation.

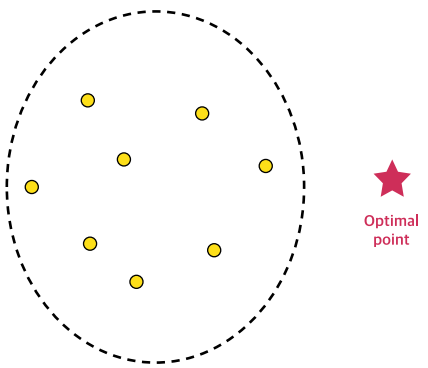
**FIGURE 2.11:** Overview of the Anticipated Mean Shift (AMS) process in MO-RV-GOMEA. It guarantees moving in the direction of improvement of the cluster. Dashed ellipses represent previous generation clusters (being the mean the blue point) and the filled ellipses represent the current generation cluster (being the mean represented with a star). The AMS is the vector connecting the old generation mean with the new generation mean.

is steered into high-quality solutions in terms of guidance values, only preserving best solutions. These include solutions between the best guidance value and 1.5 times this value. Thus, the model is able to give solutions with large deformations while preserving the energy as a metric in a multi-objective approach.

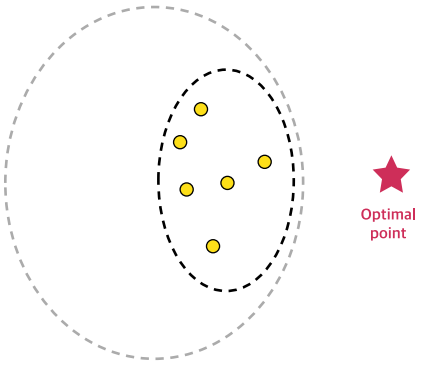
### 2.7.5 Deep learning multi-objective deformable image registration

The other method this thesis builds on is the Deep Learning Multi-Objective Deformable Image Registration (DL-MODIR) method [25] that has demonstrated to be better able to explore the trade-off space between different objectives than grid search methods. This method uses the normalized cross-correlation for the intensity, Dice similarity for the segmentation and the spatial gradient displacement for the magnitude.

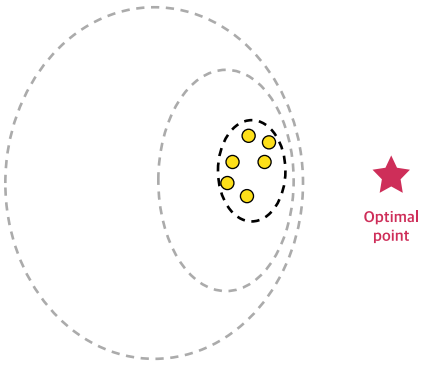
In contrast to evolutionary methods, DL is faster at inference, where results can be obtained within seconds at the expense of a previous training phase. Unsupervised DL methods contain a deep convolutional neural network model that gets



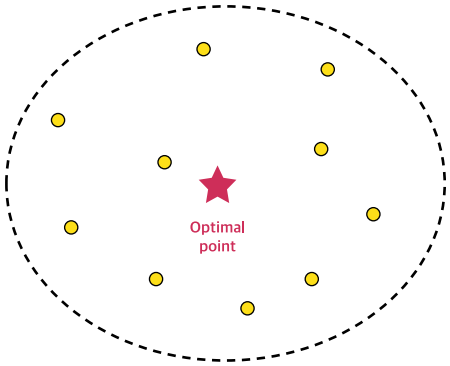

**(A)** First generation of solution collapse.




**(B)** Second generation of solution collapse.




**(C)** Third generation of solution collapse.




**(D)** Solution overshoot.

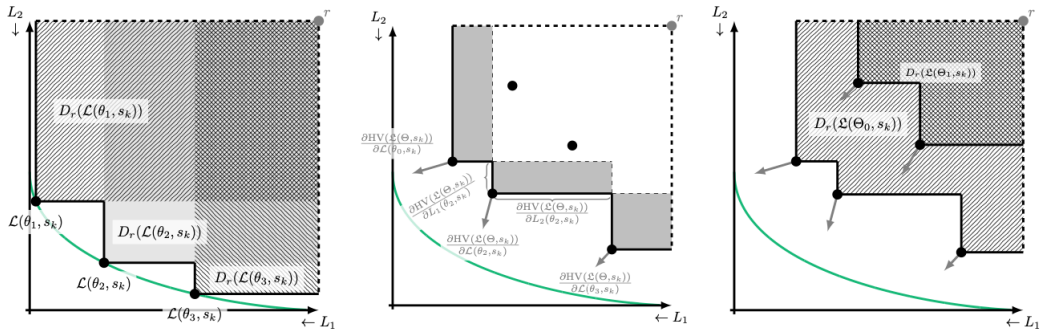
**FIGURE 2.12:** Overview of two cases where the distribution multiplier needs to be adapted to adjust the variance of the Gaussian. (a, b, c) Solution collapse. First (a), second (b), and third (c) generation solutions going to the optimal point (red star). Solutions collapse over generations, and the variance of the Gaussian needs to be increased to converge faster. (d) Overshooting in the optimal solution. Generated solutions are spread around the optimal solution. It is necessary to decrease the variance of the Gaussian to get closer to the optimal.

the source and target images as input and returns the DVF as output. The model is trained to learn the parameters that minimize the loss function. Usually, this loss function consists of a linear scalarization of registration quality metrics [7]. However, some studies [39, 42] have treated DL methods from a multi-objective point of view that return an approximated Pareto front instead of a unique solution.

One strategy to optimize a multi-objective DIR problem with DL methods consists of maximizing the hypervolume of the approximated Pareto front [44]. The method used in this thesis uses a dynamic loss function based on hypervolume maximization [18] for multi-objective optimization. Considering $k$ objectives to be optimized, each network tries to minimize a vector of losses:

$$\mathcal{L}(\theta, s_n) = [L(\theta, s_n), ..., L_k(\theta, s_n)] \tag{2.17}$$

with $S = \{s_1, ..., s_n, ..., s_{|S|}\}$ samples. Considering $p$ networks, each network $i \in \{1, ..., p\}$ corresponds to a solution of the approximated Pareto front. The goal is to optimize the model's parameters $\Theta = \{\theta_1, ..., \theta_i, ..., \theta_p\}$ for each individual




(a) Dominated subspaces (b) HV gradients (c) Domination-ranked fronts

**FIGURE 2.13:** Hypervolume maximization with a dynamic loss function. Figure from [18].

network $i$ to minimize a loss function $\mathcal{L}_i(\theta_i, s_n)$. Then, all losses for all networks can be stacked in a matrix as follows:

$$\mathfrak{L} = \begin{bmatrix} L_1^1 & L_1^2 & ... & L_1^p \\ L_2^1 & L_2^2 & ... & L_2^p \\ ... & ... & ... & ... \\ L_k^1 & L_k^2 & ... & L_k^p \end{bmatrix} \tag{2.18}$$

Considering that $\mathfrak{L}$ contains all loss functions for $p$ models with $k$ objectives, the main goal is to maximize the hypervolume of the approximated Pareto front:

$$\text{maximize} \frac{1}{|S|} \sum_{n=1}^{|S|} \text{HV}(\mathfrak{L}(\Theta, s_n)) \tag{2.19}$$

Therefore, a gradient ascent algorithm is applied to update $\Theta$ such that the hypervolume is maximized. By the chain rule decomposition Equation 2.19, as described in [18], can be converted to:

$$\frac{1}{|S|} \sum_{n=1}^{|S|} \frac{\partial \text{HV}(\mathfrak{L}(\Theta, s_n))}{\partial \mathcal{L}(\theta_i, s_n)} \frac{\partial \mathcal{L}(\theta_i, s_n)}{\partial \theta_i} \quad \forall i \in \{1, ..., p\} \tag{2.20}$$

This equation can be further decomposed for each objective such that each solution is optimized in a different direction in the objective space:

$$\frac{1}{|S|} \sum_{n=1}^{|S|} \sum_{j=1}^{k} \frac{\partial \text{HV}(\mathfrak{L}(\Theta, s_n))}{\partial L_j(\theta_i, s_n)} \frac{\partial L_j(\theta_i, s_n)}{\partial \theta_i} \quad \forall i \in \{1, ..., p\} \tag{2.21}$$

where $\frac{\partial \text{HV}(\mathfrak{L}(\Theta, s_n))}{\partial L_j(\theta_i, s_n)}$ is the hypervolume (HV) gradient for a single objective and $\frac{\partial L_j(\theta_i, s_n)}{\partial \theta_i}$ is the gradient used in gradient descent to optimize the model parameters for a single objective. This means that the more the hypervolume increases, the bigger the gradient of the loss function $\frac{\partial L_j(\theta_i, s_n)}{\partial \theta_i}$ is in the direction that reduces that loss for objective $j$ (Figure 2.13b). Therefore, this method minimizes the loss $L_j(\theta_i, s_n)$ and simultaneously maximizes the hypervolume.

Existing work [18] goes further into this method by considering domination-ranked fronts. A caveat of the described method is that strongly dominated solutions (solutions in the interior of the approximated Pareto front) have zero gradient:

$$\frac{\partial \mathrm{HV}(\mathfrak{L}(\Theta, s_n))}{\partial L_j(\theta_i, s_n)} = 0 \tag{2.22}$$

since any movement does not increase the hypervolume. To solve this problem, the paper proposes to optimize different fronts in parallel, one for each rank (Figure 2.13c). Solutions in the Pareto front correspond to rank 0. To compute rank 1, solutions from rank 0 are discarded, and the non-dominated solutions from the remaining solutions are considered. This is done iteratively until the entire set of solutions is covered. Finally, the dynamic loss is averaged over multiple samples for more stable computation.

The DL-MODIR method used in this thesis is based on the VoxelMorph architecture [7] adapted for multi-objective training where, instead, $p$ neural networks are used and the dynamic loss function from Equation 2.21 is optimized.

The method proposed by Grewal et al. [25] uses a shared encoder for all the networks and individual decoders ($p$ in total) for each solution of the approximated Pareto front (Figure 2.14).

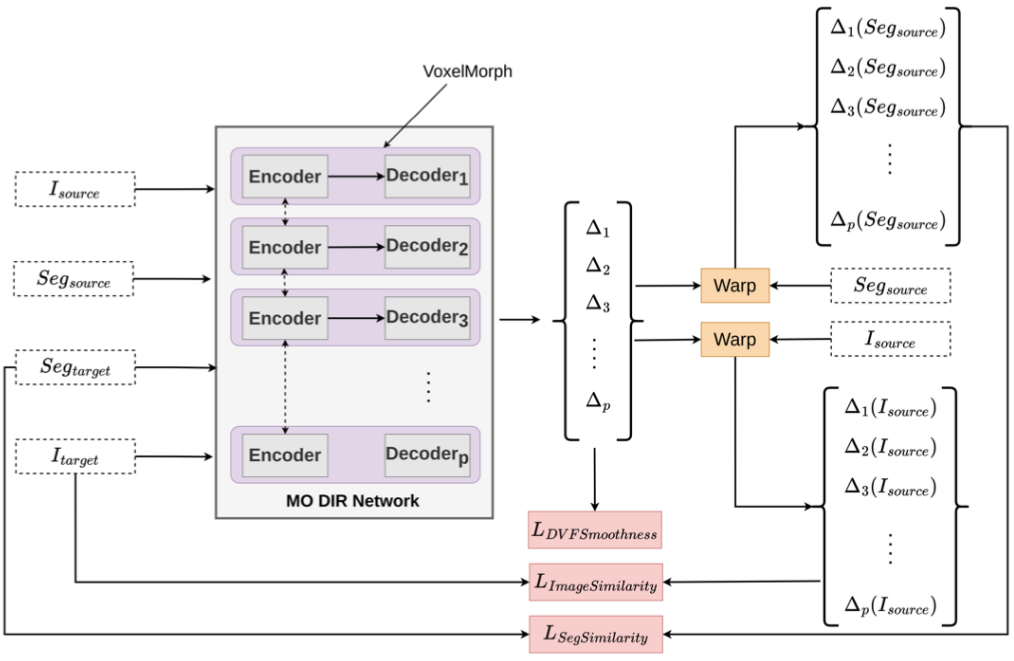



**FIGURE 2.14:** Pipeline of DL-MODIR adopted in this thesis. Figure from [25]. DL-MODIR uses the VoxelMorph deep neural network with $p$ decoders. The resulting DVFs are used to compute the transformed source image and segmentation, which are used to compute the image similarity (intensity objective) and segmentation similarity (segmentation objective). The smoothness (magnitude objective) is computed directly from the output DVF.

## 2.8 Comparison between deep learning and evolutionary algorithm optimization methods

When tackling DIR problems, it is crucial to understand the inherent trade-off between optimization quality and computational efficiency. While FEM-based methods offer highly realistic solutions, their reliance on EA demands considerable time to optimize each sample, which can rise to several hours, depending on the scan resolution and the size of the field of view. Consequently, these methods currently remain impractical for some clinical settings despite showing promise in finding realistic transformations.

In contrast, DL methods have significantly faster inference times and can return a DVF within seconds in most cases. However, this speed comes at the expense of realism, as each voxel deformation is computed independently, potentially resulting in highly unrealistic transformations. Even with the inclusion of energy metrics, DL methods can still fall short of realistic solutions. Moreover, unlike EA, these methods require an extensive training. Table 2.1 shows the differences between both methods.

The independent deformation of each voxel in the DL method allows for finer-grained deformations, compensating for the loss of information inherent in the generation of mesh grids (containing fewer nodes than voxels). However, this increased granularity often comes at the cost of realism, as finer deformations may not accurately reflect the underlying anatomical structures, being more prone to folds.

| | **DL-MODIR** [25] | **MOREA** [4] |
|---|---|---|
| **Transformation model** | End-to-end | FEM |
| **Solutions** | Multi-Objective | Multi-Objective |
| **Optimization method** | Deep Learning | Evolutionary Algorithm |
| **Training phase** | Yes | No |
| **Sample optimization** | Seconds | Hours |

**TABLE 2.1:** Comparison between both methods used in this thesis: a DL approach with an end-to-end transformation model and an EA method with a FEM.

## 2.9 Hybrid methods

Multiple DIR methods have been proposed in the literature, and the most relevant ones are described in this chapter. However, these methods must deal with a trade-off between accuracy and speed. While some methods seem to perform more fine-grained and accurate registration, they are extremely time-consuming, making them less feasible for clinical practice. For this reason, hybrid methods have the potential to be a solution. They exploit the benefits of each method and combine them into a unique pipeline that balances the trade-off.

This has been previously attempted by Huo et al. [30] with a hybrid method between an automatic thin plate splines-based and an improved Demons algorithm that outperforms these methods treated individually. As posed by the paper, Demons can accurately correct local differences, but it is time-consuming.

The thin plate splines-based method is used as a first step to perform a preliminary transformation, followed by Demons for a more fine-grained transformation (Figure 2.15). The transformed source image obtained using the thin plate spline method is given to Demons as input.

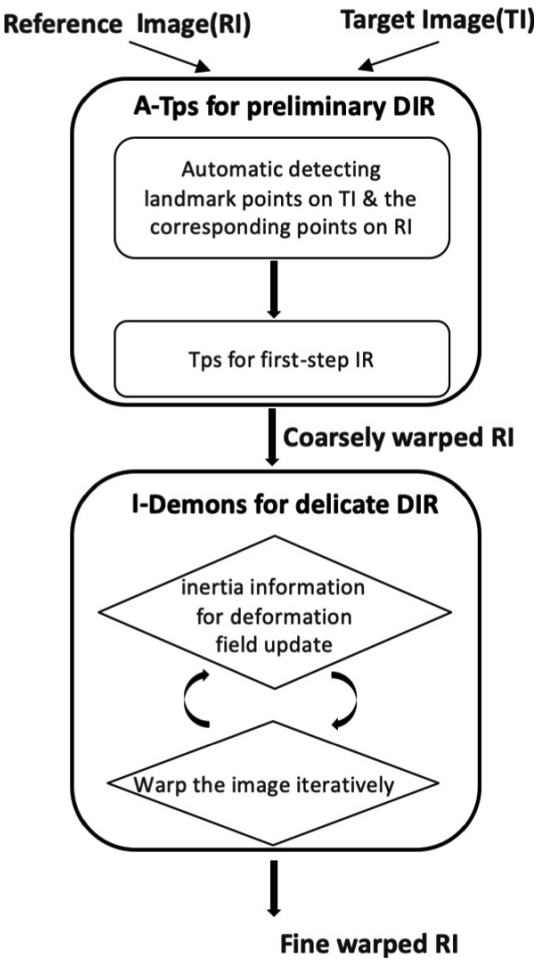



**FIGURE 2.15:** Flowchart of the proposed hybrid method with a splines-based method followed by a Demons method for a more fine-grained registration. Image from the original paper [30].

The results reveal a gap that can be filled in a promising way by combining several methods in a hybrid. Experiments in five different datasets show that the hybrid method outperforms each individually regarding accuracy and computational time, as displayed in Table 2.2. The evaluated accuracy is computed as the mean squared intensity difference between the target and the transformed source image.

However, the proposed method contains some limitations. It only works with 2D images and uses single-objective DIR. The thin plate splines-based method uses a linear scalarization consisting of landmark distance difference and a regularization term for the energy. Landmarks are automatically detected in a process that can introduce errors in quality registration metrics. Demons, instead, uses the intensity and energy in a single objective. Also, it is unclear how the deformation

| Problem | Metric | Demons | B-spline | Hybrid |
|---|---|---|---|---|
| lenag | MSD ($\cdot 10^{-4}$) | 3.746 | 3.862 | **2.419** |
| | Time (s) | 1.815 | 3.7 | **0.621** |
| rose | MSD ($\cdot 10^{-4}$) | 49 | 35 | **25** |
| | Time (s) | 8.893 | 44.5 | **5.778** |
| brainT1 | MSD ($\cdot 10^{-4}$) | 0.217 | 0.755 | **0.62** |
| | Time (s) | 2.267 | 11.647 | **1.301** |
| sunflower | MSD ($\cdot 10^{-4}$) | 0.361 | 0.8 | **0.434** |
| | Time (s) | 2.43 | 18.916 | **2.11** |
| brainT2 | MSD ($\cdot 10^{-4}$) | 0.194 | 0.651 | **0.114** |
| | Time (s) | 1.833 | 6.45 | **1.002** |

**TABLE 2.2:** Results from the hybrid method paper [30] comparing the Mean Square Difference (MSD) and the execution time in seconds for the two methods individually and the hybrid method. In bold is displayed the best model for each problem-metric pair. The hybrid method outperforms both methods individually in every problem in terms of accuracy and time.

of both methods is combined into a single DVF. To the best of our knowledge, this is the only work that exists combining fundamentally different DIR methods into a hybrid. This thesis is dedicated to further analyzing the potential of hybrid methods in the field of DIR by combining a MOREA and DL-MODIR in a hybrid method.

Chapter 3

# Methods

This chapter explores the developed hybrid method, its main components, and their interconnections. We begin by introducing the three components of the hybrid method. We then describe the developed approach and its application with a synthetic dataset. We also demonstrate its behavior on a clinical dataset for cervical cancer RT planning by performing sanity checks of the hybrid method on this dataset. Finally, we present some preliminary experiments and list the chosen hyperparameters.

## 3.1 Hybrid method

This thesis aims to develop a hybrid method that combines two existing methods, MOREA (Section 2.7.4) and DL-MODIR (Section 2.7.5). By leveraging the fast output of DL-MODIR at inference to smartly initialize MOREA's population for a warm start, we enhance the efficiency and accuracy of the registration process. As a FEM-based transformation model, MOREA can perform more realistic registrations than DL-MODIR, making it less prone to folds linked to unrealistic deformations. The goal is for DL-MODIR to provide MOREA with sufficient information to start the process with an advantage over the original MOREA version, ideally outperforming both methods separately.

The hybrid method contains three components:

- **DIR Problem**. It stores the pre-processed data in an object for reading and writing it easier between the different methods.
- **DL-MODIR method**. It contains the DL-MODIR method described in Section 2.7.5 with some adaptations.
- **MOREA method**. It contains the MOREA method described in Section 2.7.4 with some adaptations.

These three elements connect in a pipeline divided into two phases: training and inference. Figure 3.1 displays how these three components are connected and which steps occur in the training and inference phases. DL-MODIR updates its neural network parameters during training, using 80% of the dataset problems. This step is only necessary once, since the goal is for the learned parameters to generalize to unseen problems. The learned neural network is then used for inference and evaluated on the remaining 20% of the dataset. Subsequently, DL-MODIR predicts the inverse DVF occurring in each DIR Problem and stores it to be used by MOREA later for a warm start in the hybrid method.

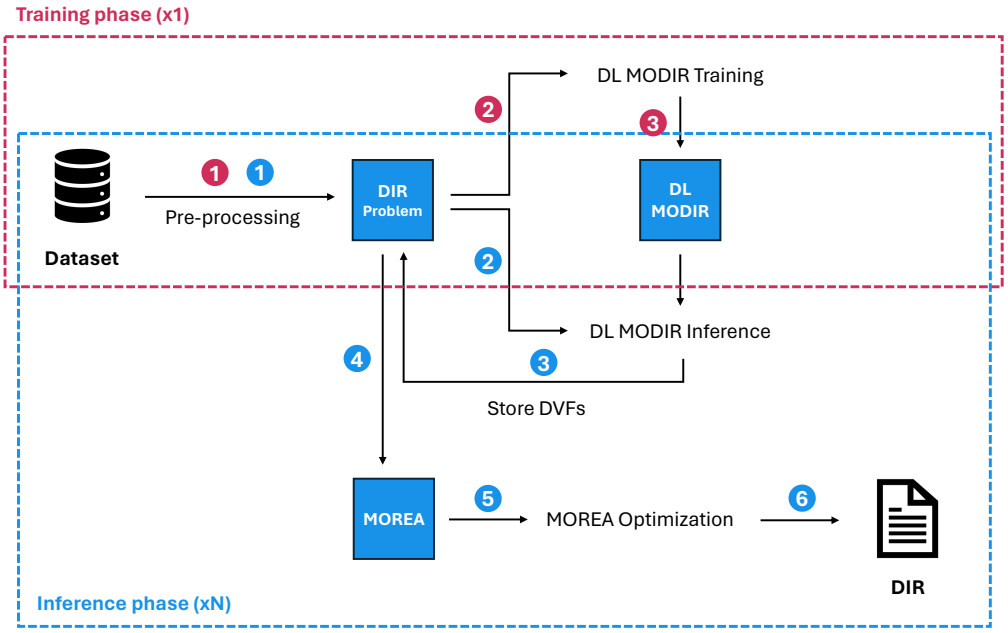



**FIGURE 3.1:** The pipeline of the hybrid method shows the connection between the DIR Problem, DL-MODIR, and MOREA. The process is divided into two phases: training and inference. MOREA is utilized only for inference optimization while DL-MODIR works in the two phases.

The following sections elaborate on the three main components, describing their tasks and adaptations in the developed hybrid method. For simplicity, we call DVF to the inverse DVF. All the DVFs represented and mentioned in the methods and results chapters correspond to inverse DVFs.

### 3.1.1 DIR Problem

A DIR problem is an object that contains the necessary information for the registration. Every DIR Problem contains the following information:

- Source image: a three-dimensional array with the image volume.
- Target image: a three-dimensional array with the image volume to be achieved by transforming the source image.
- List of objects: each object (organ) is a distinguished and segmented element on the image. Each object contains:
  - Source mask: a three-dimensional binary mask of the specific object in the source image given by the user. It has the same shape as the source image.
  - Target mask: a three-dimensional binary mask of the specific object in the target image given by the user. It has the same shape as the target image.

  - Contour pairs: contours for each mask, source and target, automatically computed for each given mask with a binary erosion operation.
  - Distance map: distance between each image voxel and the closest contour point for that object. It is measured automatically for the given masks and computed contour pairs. MOREA uses it to calculate the guidance objective.

- Landmarks: set of anatomically relevant points coupling equal landmarks in each image. These points are given by the user.

Every DIR Problem object is readable by both methods, DL-MODIR and MOREA. In this thesis, DL-MODIR is adapted explicitly for loading a DIR Problem object and using it to train and evaluate the neural network. Figure 3.2 displays an object diagram with the components of a DIR Problem object and their connection.

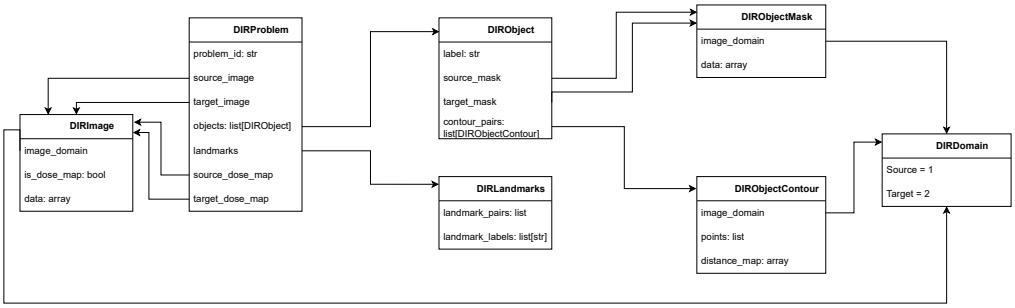



**FIGURE 3.2:** Object diagram of a DIR Problem with the connection between its components in an object-oriented programming design.

### 3.1.2 DL-MODIR

This component contains DL-MODIR with some adaptations to include it in the hybrid method and interact with MOREA. Compared to MOREA, some requirements constrain the format of the given data. For instance, the input size must be fixed with square slices (in-plane) of unrestricted depth. DL-MODIR handles the depth variability by looking at a window of slices with a fixed size, manually defined as the maximum depth $d_{max}$. Different slices are selected depending on the phase:

- Training: a random slice is selected, and the consecutive slices, with a length $d_{max}$, are taken as the input. As a result, the neural network receives a random fraction of the image. Hence, the generalization capabilities are enhanced by selecting different fractions at different epochs.

- Inference: the entire image is evaluated in independent fractions of $d_{max}$ consecutive slices. The final DVF is the result of connecting the partial DVFs and averaging over the shared slices.

DL-MODIR, compared to MOREA, has two distinguished phases: training and inference. During training, the deep neural network updates its parameters for the fraction of the training dataset with a gradient descent algorithm using a dynamic

loss function (Section 2.7.5). It simultaneously minimizes the loss for all objectives –defined by the registration quality metrics–, and maximizes the hypervolume of the approximated Pareto front. Training takes place once, aiming to learn a function that, when evaluated on data out of the learned distribution, is able to generalize. The learned neural network is stored and used for inference. A total of $p$ DVFs are generated for each given DIR Problem, one for each solution in the approximated Pareto front.

### 3.1.3 MOREA

MOREA is the last method in the pipeline. It is responsible for enhancing DL-MODIR, ideally improving their realism. MOREA works in a three-step process: sweep generation, mesh creation, and optimization.

During sweep generation, the hyperparameters for the optimization run are specified. These include the mesh generation strategy, initialization method, and number of generations, among others. If a list of different values is given for one or multiple hyperparameters, all combinations of hyperparameters –as grid search– are evaluated. The sweep generator provides MOREA with the necessary information to read and optimize the DIR problem.

Later, MOREA generates a mesh for each sweep by considering the defined hyperparameters in the previous step. The mesh creation step is adapted to support alternative node placement strategies. While the original method takes the entire object to place nodes on its surface, the developed adaptation only places them around a specific object of interest (or organ). In the case of the clinical dataset, it is placed around the bladder. However, only a fraction $r_s$ of them are placed on the surface of the defined object, and the rest are randomly placed around the image using a Sobol sequence to ensure a more uniform distribution.

In the last step, optimization, MOREA reads the generated mesh grid and the sweep configuration to generate the initial population for the optimization problem. With EAs, the model aims to converge into a larger set of more optimal solutions that satisfy the three different objectives: magnitude, intensity, and guidance. The most significant adaptation to developing the hybrid method is the smart population initialization in MOREA, which permits a warm start.

## 3.2 Developed solution: MOREA warm start

EAs need an initial population to start optimizing from. Although a random initialization is often performed, we think the closer we start from the optimal solution (or approximated Pareto front in multi-objective optimization), the faster the model can converge.

### 3.2.1 Gaussian attractors initialization

The default population initialization performed by MOREA takes the generated mesh and applies a fixed number of randomly placed radial-based attractors with a weight sampled from a uniform distribution (between 0 and 1). These attractors move mesh nodes towards them. Each solution from the initial population has

attractors with different locations and weights, ensuring diversity in the initial population. Mesh nodes are disrupted in multiple steps to ensure that the movement is not entirely restricted when a DVF generates folds. In this case, only a partial movement is done until the fold restricts the movement of the mesh node.

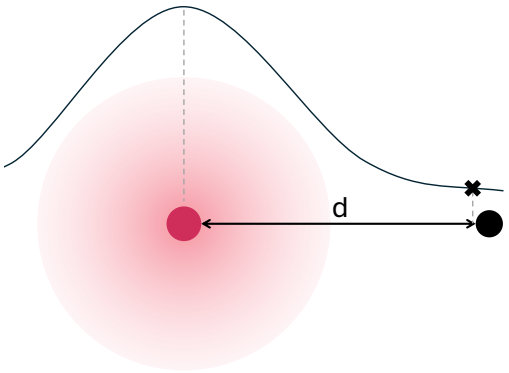



**FIGURE 3.3:** Gaussian attractors intensity. Distance $d$ defines the attraction intensity $I_a$, sampling from a Gaussian distribution where the intensity is in the cross of the Gaussian. Hence, the attracting intensity decays with the distance.

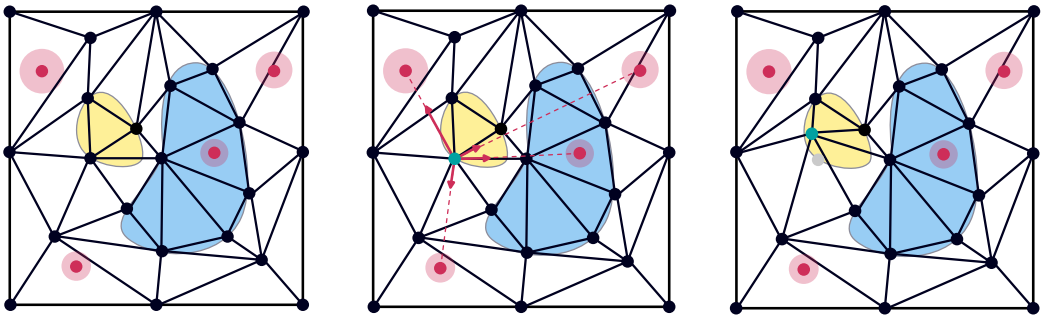
As displayed in Figure 3.3, the distance $d$ between the node and the attractor defines the attracting intensity by sampling from a Gaussian distribution. This distribution has a mean zero in the center of the attractor and a standard deviation equivalent to the maximum distance space divided by 16. The maximum distance space is computed as the image volume's diagonal, equivalent to $\sqrt{3}$ when the mesh node coordinates are normalized in the range $[0, 1]$. The attraction intensity $I_a$ results from computing the value $d$ in the Gaussian distribution, and it is defined as follows:

$$I_a = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.1}$$

where the mean is $\mu = 0$, the standard deviation in the normalized space is $\sigma = \sqrt{3}$, and $x = d$. Hence, mesh nodes closer to the attractor (lower $x$) move more towards the attractor than further nodes, where the attraction intensity gets attenuated with a Gaussian decay. The attractor weight $w_a$ acts as a multiplier for the attraction intensity, with lower weights producing less attraction than higher weights. Therefore, the magnitude of the attraction force $F_a$ that a Gaussian attractor produces in a mesh node is defined as:

$$F_a = I_a \cdot w_a \tag{3.2}$$

For each mesh node, the algorithm computes the attraction force all attractors exert on it and divides it by the number of attractors.



**(A)** Gaussian attractors.

**(B)** Forces applied to a mesh node by the attractors.

**(C)** Movement of the mesh node.

**FIGURE 3.4:** MOREA initialization with Gaussian attractors. Attractors (red dots), with different weights (red shadow), exert attraction forces to each mesh node with a Gaussian distribution intensity. The resulting movement comes from summing all the attraction forces that the attractors exert on the mesh node and normalizing by the number of attractors.

### 3.2.2 Warm start with DVF solutions

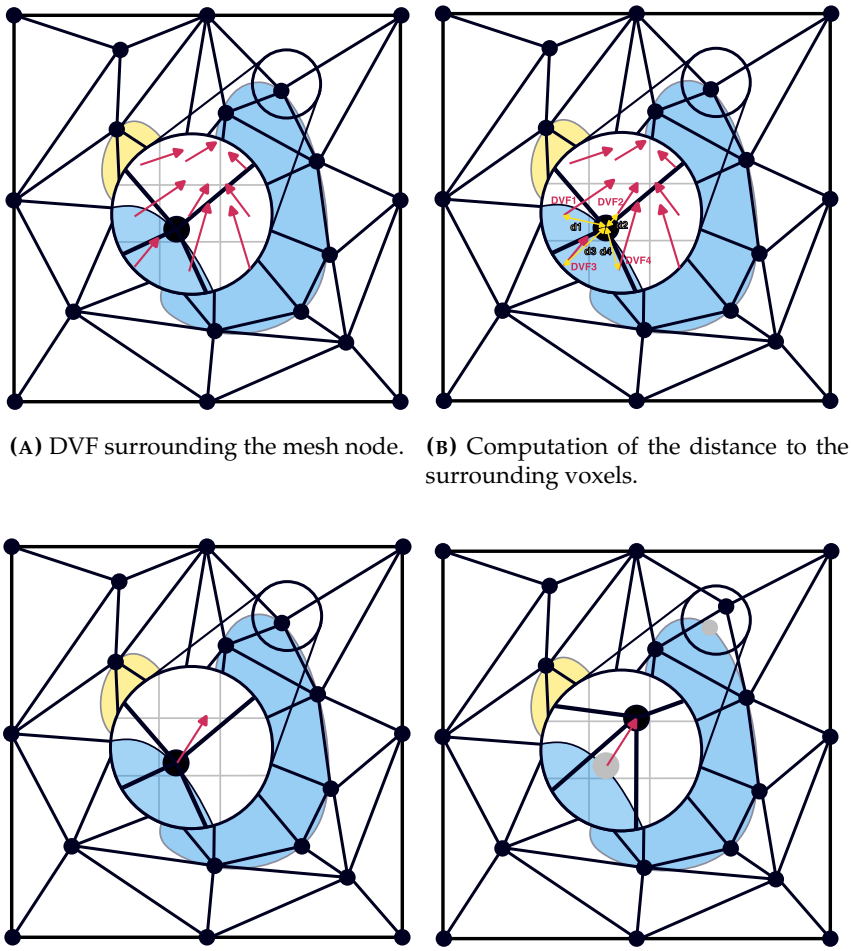
In the developed hybrid method, the population initialization based on the Gaussian attractors performed by MOREA is changed to apply the DVFs from DL-MODIR. MOREA leverages DL-MODIR solutions to perform a warm start with its information. For every solution forming the initial population, a DVF from DL-MODIR is applied to the mesh nodes of the source image. DL-MODIR gives $p$ solutions, each containing a DVF in a different location of the approximated Pareto front. Considering that the initial population –with a population size $S_p$– needs to be filled and $S_p > p$, the hybrid method must generate multiple solutions from each DL-MODIR solution. For this reason, for each DL-MODIR solution, the DVF is given in different intensities to fill the entire initial population. Hence, only $p$ solutions from $S_p$ contain the complete DVFs from DL-MODIR; the rest are attenuations of these $p$ solutions.

In total, $\lfloor S_p/p \rfloor$ intensities $I$ are considered for each DL-MODIR solution. These intensities range from $[\frac{1}{\lfloor S_p/p \rfloor}, 1]$ where $I = 0$ would mean no transformation, and $I = 1$ means that the complete DVF from DL-MODIR is applied to the mesh grid. Intermediate values scale each displacement vector of the DVF by their intensity. For example, if $I = 0.5$, each displacement vector in the DVF is half the length of the original.

Given a population size $S_p = 600$ and $p = 15$ solutions in DL-MODIR, there are 40 solutions in MOREA for each DL-MODIR solution in the initial population. Each solution has a different intensity, chosen from 40 equally spaced intensities ranging from 0.025 to 1.

MOREA updates mesh nodes sequentially. As this is prone to folds, the mesh node is updated in a multiple-step process, where $N_s$ steps are done. For each step $n_s \in \{1, ..., N_s\}$, we update all mesh nodes with an intensity of $\frac{\text{DVF} \cdot I}{N_s}$ in the direction of the DVF considering the initial position of the mesh point. If we considered the new location of the mesh point after each update step, the original mesh node would not reach the location its DVF is pointing to and the DVF would change

along the multi-step process, recomputing the interpolated DVF at each step new location. The update order is done randomly at every step $n_s$ to avoid that some areas of the image are more deformed than others.




**(A)** DVF surrounding the mesh node.

**(B)** Computation of the distance to the surrounding voxels.

**(C)** Resulting $\text{DVF}_n$ of the mesh node.

**(D)** Movement of the mesh node.

**FIGURE 3.5:** Application of a DVF to a mesh grid, illustrating the process at a single mesh node. The resulting node's DVF ($\text{DVF}_n$) is derived from its neighboring voxels' interpolation. The distance to the center of these surrounding voxels is used to compute a weighted average, determining the final deformation applied to the mesh node.

A representation of how this method works is displayed in Figure 3.5. For a mesh node, some displacement vectors surround it, each value corresponding to the displacement at the center of each voxel. However, mesh nodes do not need to be in the center of a voxel, requiring some interpolation to compute the DVF in its position. In two dimensions, the four voxels surrounding the mesh node are considered to define the final direction of the DVF. First, the distance between the mesh node and the center of each voxel is computed; these distances $d_1, d_2, d_3, d_4$ are used to calculate a weighted average of the DVF. At the end, the resulting $\text{DVF}_n$ for the mesh node in Figure 3.5 is computed as:

$$\text{DVF}_n = \frac{\text{DVF}_1 \cdot d_1 + \text{DVF}_2 \cdot d_2 + \text{DVF}_3 \cdot d_3 + \text{DVF}_4 \cdot d_4}{4} \tag{3.3}$$

The figure simplifies the developed method to a two-dimensional representation, but the datasets contain three-dimensional volumes. Hence, the eight surrounding pixels must be considered to compute the DVF interpolation instead of four.

As MOREA has a dual-dynamic model, there is a mesh grid for the target image, and the DVF is computed from the two mesh grids. The target mesh nodes in the initial population of the developed hybrid method are disrupted with Gaussian attractors.

### 3.2.3 Half-hybrid method

We developed an alternative method that combines the Gaussian attractor initialization with the DVF initialization and we called it half-hybrid. This method aims to have the potential of the DVF initialization by generating some solutions from DL-MODIR DVFs while saving half of the initial population solutions for the Gaussian attractors' initialization. Since only half of the solutions are given for the DVF initialization, less intensities are generated from each DL-MODIR solution. We expect the developed half-hybrid method always to be comparable to or better than MOREA.

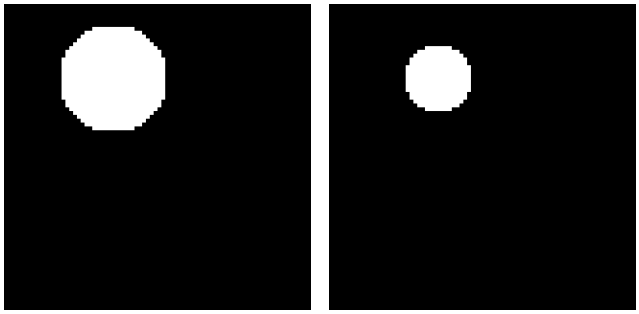
### 3.2.4 Sanity checks on the developed hybrid method

Some sanity checks are done to test the correct implementation and desired behavior of the developed hybrid method. These sanity checks are first evaluated on a synthetic dataset with a toy problem, a shrinking sphere. Each dataset problem consists of a pair of images with size $64 \times 80 \times 80$ (depth, height, and width), where the source image contains a sphere with a radius between 20% and 40% of the smallest image dimension (64). The center of the sphere is located in a random position of the image, guaranteeing a minimum space from the sphere's surface to the edge of the image of 5%. The target image consists of a shrink of the source sphere, with the same center coordinates but reducing its radius between 20% and 40% by sampling from a uniform distribution. The synthetic dataset consists of 200 problems with paired images; 80% of them form the training split, and the remaining 20% the evaluation split. Figure 3.6 displays an example of a problem from the synthetic dataset.

The first step consists of training DL-MODIR for this task. The neural network is trained for 15,000 iterations. Since the synthetic dataset contains binary images, only black and white voxels, the intensity objective is equivalent to the segmentation objective. For this reason, this problem is treated as bi-objective, only evaluating the intensity objective –with the normalized cross-correlation–, and the magnitude objective –with the spatial gradient loss–.

After training, DL-MODIR is expected to be able to generalize to unseen problems. This means that the model should be able to register images for spheres at different positions and sizes. The visualization of the DVFs for the solutions of DL-MODIR (Figure 3.7) permits a better understanding of how this method works. Later, the resulting DVFs are applied to the mesh grid generated by MOREA.

To better understand DVFs, Figure 3.8 shows the transformed source image after applying the 15 solutions from DL-MODIR and compares it to the target image.



**(A)** Source image. **(B)** Target image.

**FIGURE 3.6:** An instance of a source and a target image from the synthetic dataset. The target image contains a smaller sphere than the source image, with its center in the same coordinates.

Despite some solutions are able to achieve a transformed source image that accurately matches the target image, in some cases the deformation is incomplete or non-existent, suggesting that some explored regions of the approximated Pareto do not perform a complete deformation.

To ensure the hybrid method initialization works as expected and that mesh nodes move in the direction of the DVF at their positions, we visualize both the original mesh grid and the disrupted mesh with the DVFs applied. Figure 3.9 shows the movement of the mesh nodes after applying all DVFs with $I = 1$. Blue nodes represent the original mesh grid generated by MOREA, and red nodes are the disrupted ones after applying the DVFs from DL-MODIR at initialization. Yellow arrows represent the displacement vector of the DVF in the position of the original mesh nodes, where the arrow's tip indicates the new position each mesh node should reach. If the red point is halfway along the yellow arrow, constraints such as folds are limiting its full movement. In Figure 3.9 all nodes move in the direction of the DVF, indicating a correct implementation of the DVFs initialization, and almost all nodes reach the tip of the arrow, pointing that almost no folds occur when applying the DVFs.

Since MOREA uses a dual-dynamic model, the target mesh also needs to be disrupted. To ensure variability in the population and due to the intrinsic properties of the computation of the DVF from the mesh deformation, the Gaussian attractors' initialization used in MOREA is applied to the target mesh. Figure 3.10 shows the initialization of one solution with the different attractors. Gaussian attractors, represented with yellow dots, exert an attraction force to the mesh grid. In the figure shown, the attractors of the five subsequent slices in front of and behind the represented slice are displayed. The distance of the mesh node to the attractors affects the attraction intensity exerted. In this image, apart from the 2-dimensional attraction in the shown plane, the distance in the z-axis, perpendicular to the plane, must also be considered for disrupting the mesh node. In Figure 3.10 most of the nodes appear to move to the bottom of the image, as there are more attractors in that region.

Once it is checked that DVFs with $I = 1$ work as expected, it is necessary to evaluate the application of different intensities and how the transformed source images from the initial population look after applying the DVFs. DVF 2 from Figure 3.9 is

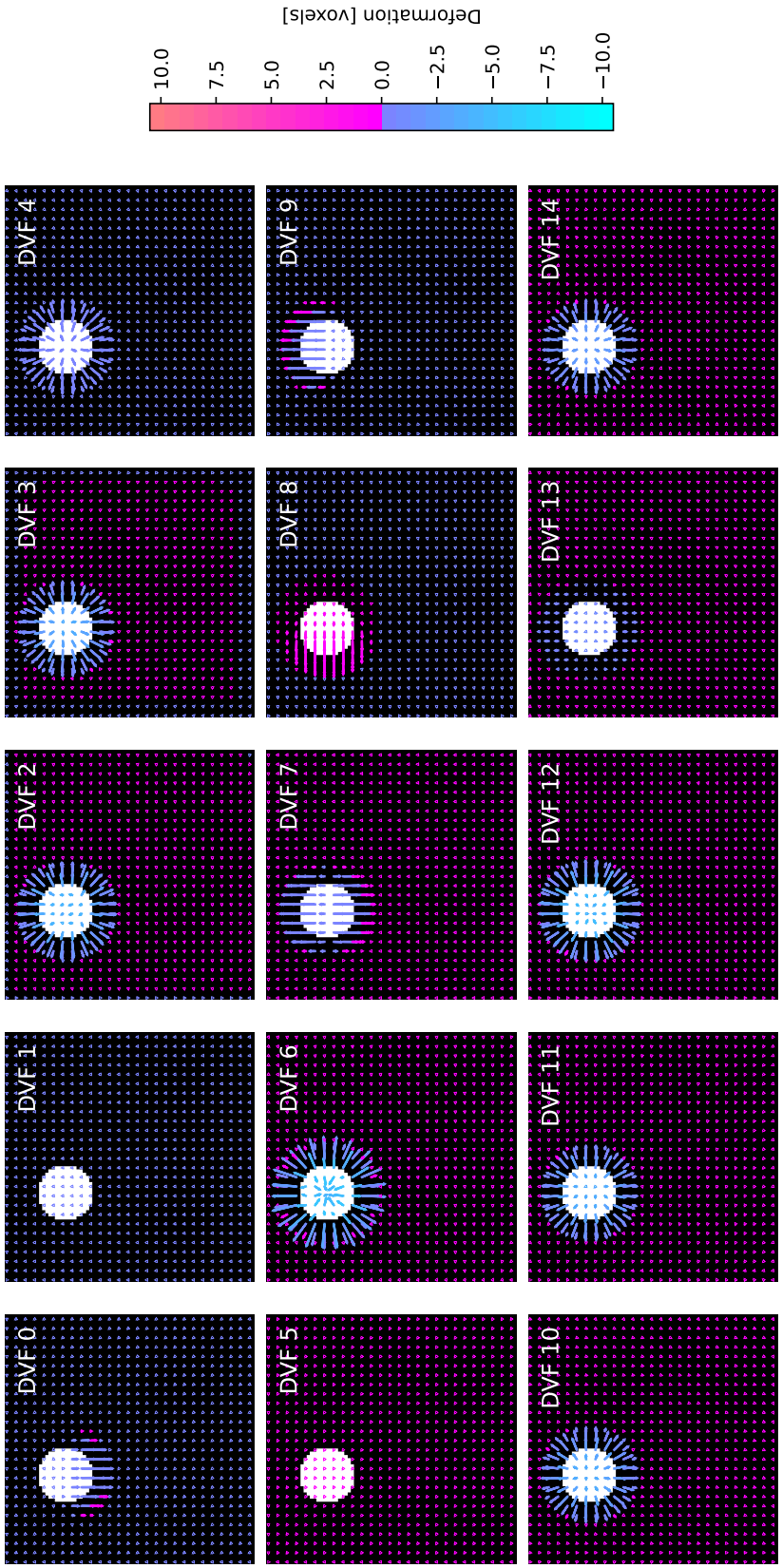



**FIGURE 3.7:** DL-MODIR DVF solutions and target image in the synthetic dataset. Each solution explores a different region of the objective space, hence different DVFs are obtained for the same problem.

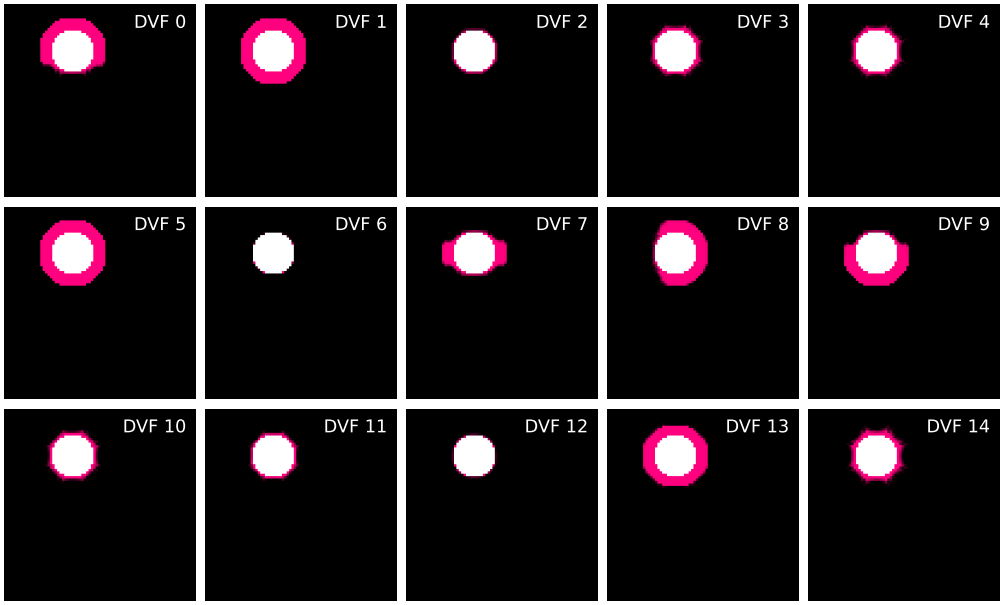



**FIGURE 3.8:** Transformed source image with the 15 DVFs from DL-MODIR. In magenta is represented the transformed source image and in green the target image. White voxels represent an overlap between the transformed source and target images.

selected for analyzing the different intensities. Figure 3.11 displays the transformed source image registered in the target image. The transformed source image compresses more when higher DVF intensities are applied. At $I = 1$, the transformed source image is very close to the target image, but the limited granularity of the mesh grid produces some imperfections in the boundaries of the sphere.

## 3.3 Cervical cancer radiation therapy planning

The synthetic dataset permits to test the developed hybrid method, but the main goal of this thesis is to evaluate it in a real-case scenario, specifically to solve the problem of cervical cancer RT planning.

### 3.3.1 Dataset

The hybrid method is trained and evaluated on a clinical dataset containing abdominal CT scans of 75 patients with IDs ranging from 02 to 76. Each patient contains a source and a target image, the source corresponding to a full bladder and the target to an empty bladder with an image spacing of 3 mm. In addition to the pair of images, the multiple organ segmentation and some landmarks for both images are also included as metadata.

The dataset is split in two, 80% of the dataset used for training the DL-MODIR neural network (patients 02 to 61) and the remaining 20% used for evaluating the developed hybrid method. The results shown in this thesis correspond to the 20% of patients used for evaluation (patients 62 to 76).

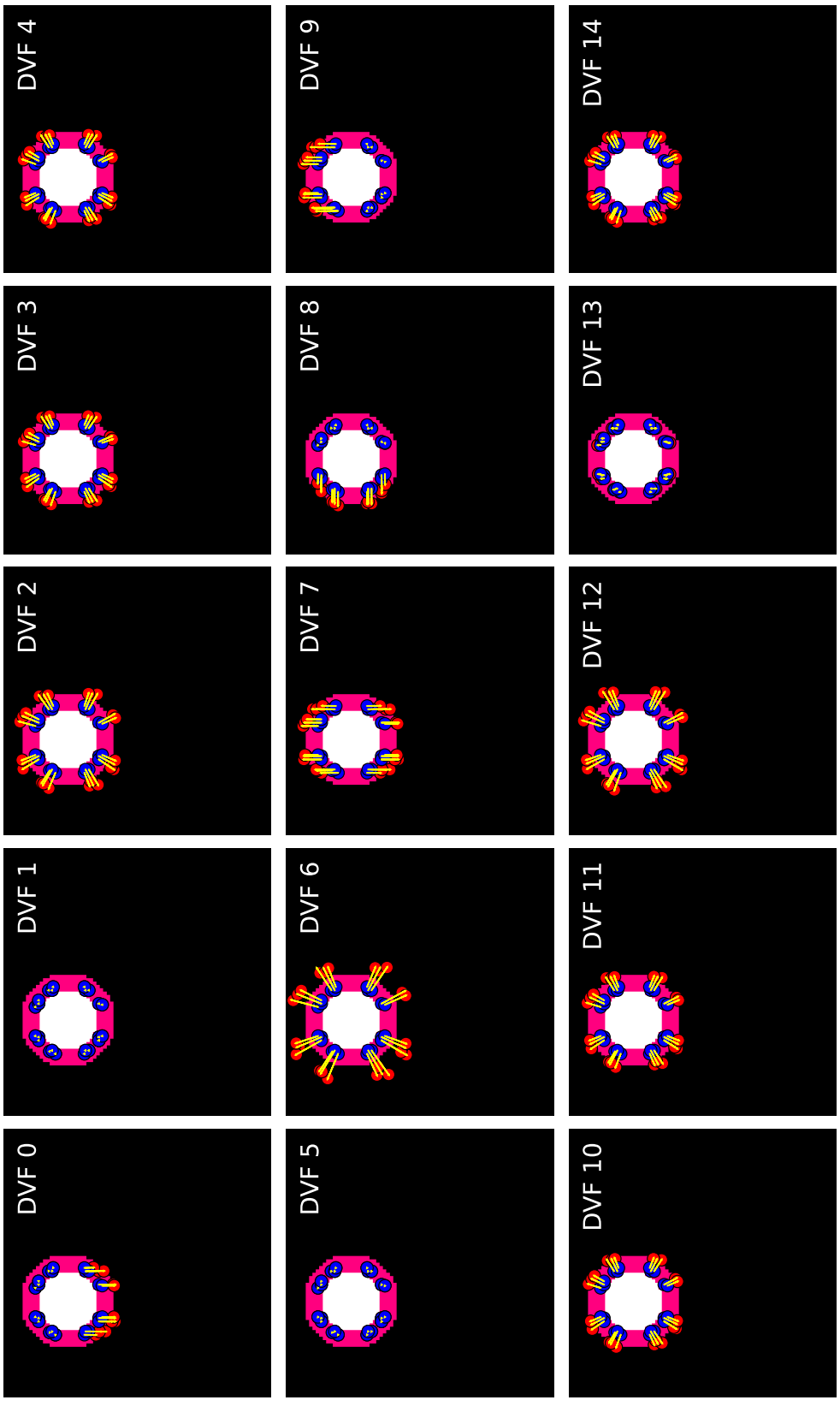



**FIGURE 3.9:** Update of the source mesh grid with the DVFs from DL-MODIR in the synthetic dataset. Blue points represent the original mesh grid, and red points represent the mesh grid disrupted with the DVFs. Yellow arrows indicate the displacement vector at the mesh node location, with the tip pointing the position the disrupted mesh node should reach. If a disrupted mesh node does not reach the end of the tip, it means that constraints are limiting the complete movement.

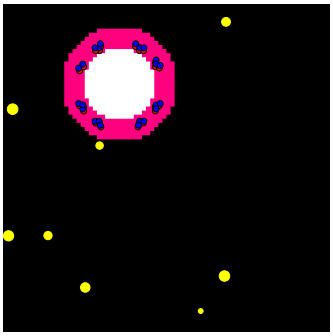


**FIGURE 3.10:** Update of the target image mesh grid with the Gaussian attractors. Blue points represent the original mesh grid, and red points represent the mesh grid disrupted by the attraction forces. Yellow points represent the Gaussian attractors, with the size indicating the weight of the attractor. Attractors with a margin of five slices are represented but the z-axis distance does not affect the plotted size.

### 3.3.2 Data pre-processing

Images from the raw dataset are not aligned nor focused on the region of interest, requiring some pre-processing to normalize data. Consequently, each patient operates the following steps:

1. **Bones rigid registration:** both CT scans are taken at different times and conditions. With rigid registration, both images are aligned, considering the bones as a guide as they are the only static organs that do not change their shape unless there is a fracture. The computed rigid registration is also applied to the segmentation masks and landmark coordinates.

2. **Bladder cropping:** this thesis focuses on cervical cancer images with a particular focus on the bladder and the deformation it exerts on the surrounding organs when emptying. Images are cropped by taking the minimum and maximum coordinates that the bladder occupy in the two images and adding some padding to study the surroundings. The padding is set to 24 mm (8 voxels) in each side of every dimension. As occurred with the rigid registration, cropping also changes the coordinates of the segmentations and landmarks, which need to be normalized to the image's new bounds.

3. **Target in-plane size:** since DL-MODIR is based on a neural network model, the input must have fixed dimensions. However, each image contains different-sized bladders, which lead to different-sized cropped images. A margin is added, filling the remaining voxels in black to reach the neural network in-plane size (width-height hyperplane) set at $96 \times 96$. The depth is set to a minimum value defined as the maximum depth $d_{max}$, set at 48. If this value is exceeded, the image is left with the same depth size, and DL-MODIR computes its DVF in fragments, as mentioned in Section 3.1.2. On the other

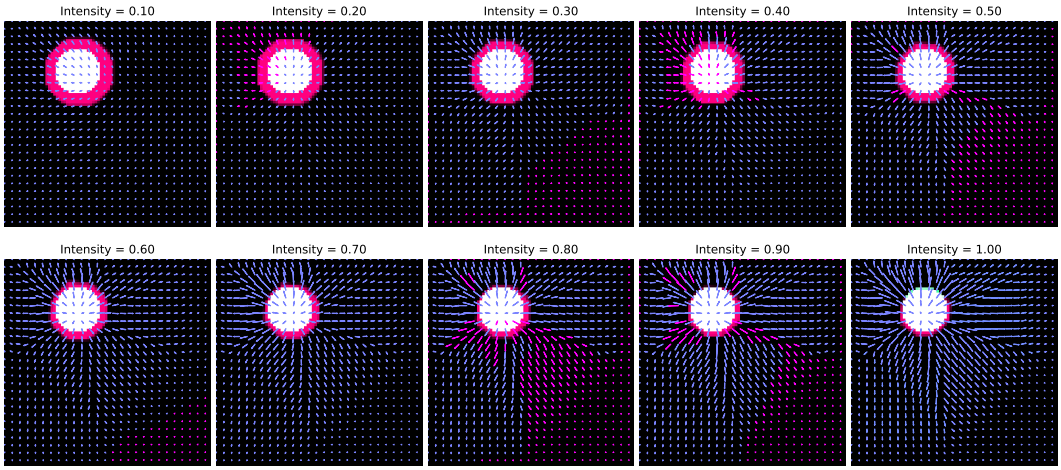



**FIGURE 3.11:** Transformed source image (magenta) and target image (green) after applying a DL-MODIR DVF with different intensities, representing different solutions of the initial population of the hybrid method.

hand, if the value is lower than $d_{max}$, it is also filled with a black margin until reaching a depth of 48. Hence, every image has the same in-plane size ($96 \times 96$) and different depths with a minimum value of 48.

After pre-processing and storing the data for each patient, both methods can directly optimize it.

### 3.3.3 Difference with the original MOREA

In some aspects, the MOREA method used in this thesis differs from the MOREA presented in the original paper [4].

In the first instance, although MOREA does not need a specific size, the pre-processed images with black margins are given for a fair comparison. Nonetheless, MOREA only places the mesh grid in the image area, neglecting the black margins, but takes them into account for computing the different registration quality metrics. Additionally, to overcome GPU restrictions in DL-MODIR, the spacing of the image is increased from 1.5mm to 3mm, resulting in a downsampled image with half resolution compared to the images optimized in the original MOREA paper.

The last difference refers to the biomechanical properties of the organs. The original MOREA considers heterogeneous elasticities, different for each organ, with bones being more difficult to deform than soft organs such as the bladder or the bowel. In the version of MOREA used in this work, all organs have the same elasticity, so bones are as difficult to deform as other organs, resulting in more unrealistic deformations.

These reasons, which may initially make the presented results for MOREA seem less favorable than the ones in the original paper [4], permits to adapt it for the hybrid method under the available GPU conditions and restrictions that DL-MODIR poses when treating images.

### 3.3.4 DL-MODIR training

First, to evaluate the hybrid method, it is necessary to train DL-MODIR on the training split. As a multi-objective problem, three registration quality metrics are evaluated, one for each objective. The intensity is measured with the normalized cross-correlation loss, the magnitude with the spatial gradient loss, and the segmentation with the dice similarity.

Figure 3.12 shows the loss evolution for each of the 15 solutions of the trained DL-MODIR along 150,000 iterations. The intensity (NCCLoss) is left stable or decreases, but the magnitude (Spatial Gradient Loss 3D) starts at low values (close to 0) and increases along iterations. Initially, the solutions show no deformation. However, as the model begins to learn from solutions with low segmentation and intensity losses, the magnitude of deformation increases because significant changes are necessary to improve these two objectives. For the segmentation, (Seg. Similarity Loss) most solutions decrease its loss, meaning that the organs of the transformed source mask match better with the target mask.

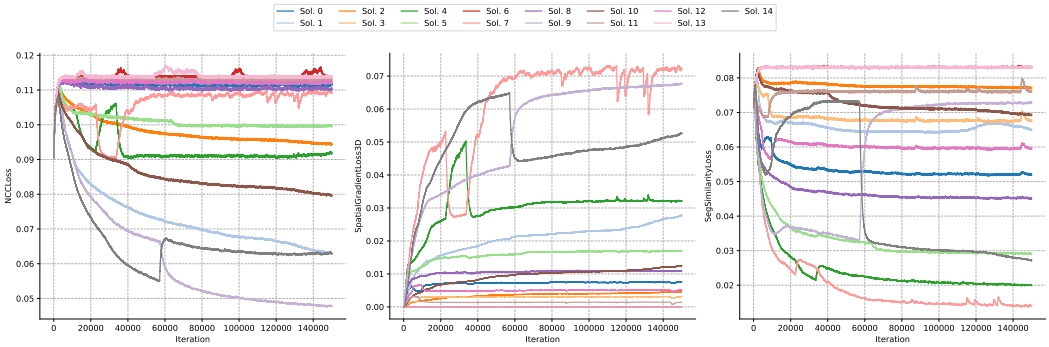



**FIGURE 3.12:** DL-MODIR train losses for the 15 solutions in three objectives: intensity measured with the NCCLoss, magnitude measured with the Spatial Gradient Loss 3D, and segmentation measured with the Seg. Similarity Loss.

After training, the model is evaluated on a separate set of patients to study the ability to generalize to unseen patients. An evaluation is performed every 3,000 iterations on the evaluation split, the same ones they are later evaluated with the hybrid method. Looking at Figure 3.13 the validation losses show a similar tendency and absolute values than the training losses indicating that the model is able to generalize to unseen patients.

DVFs returned by DL-MODIR are displayed in Figure 3.14, where different deformations are performed to explore various regions of the objective space. Remarkably, some DVFs (2, 4, 5, and 10) contain almost no deformation in one axis where displacement vectors are pointing vertically, producing a flat DVF.

The transformed source contours using DL-MODIR DVFs are included in the Appendix (Figure A.1), where some DVFs can deform the organ contours to match closely the contours of the bladder and the bowel in the target image, including some of the flat DVFs. This event suggests that a low distance between the transformed source contours and the target contours can be achieved even with unrealistic deformations.

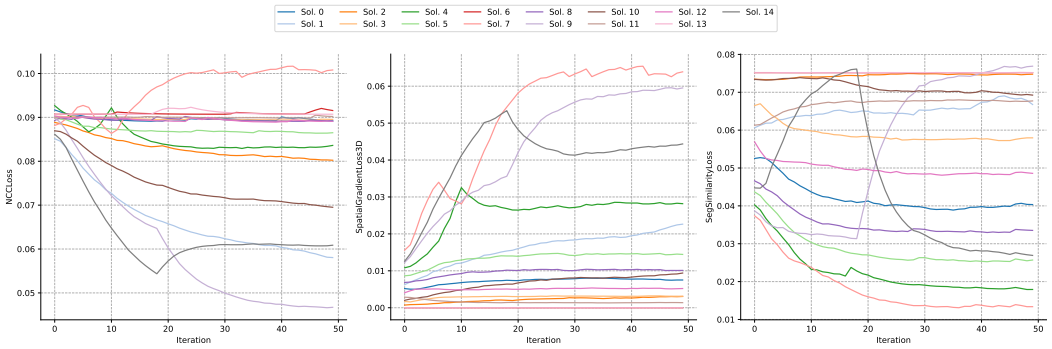



**FIGURE 3.13:** DL-MODIR validation losses for the 15 solutions in three objectives: intensity measured with the NCCLoss, magnitude measured with the Spatial Gradient Loss 3D, and segmentation measured with the Seg. Similarity Loss.

### 3.3.5 Mesh of the initial population

To confirm that the developed method works equally well on the clinical dataset as on the synthetic dataset, we also evaluate the movement of the nodes after initialization with solutions that have $I = 1$, where the complete DVF from DL-MODIR is applied. Figure 3.15 displays the nodes on the source mesh and their deformation after applying the DVFs. All nodes move in the direction of the DVF arrow; however, some do not reach the tip of the arrow. This is the case where DL-MODIR outputs a large deformation that yields folds, which is more present in this dataset than in the synthetic one.

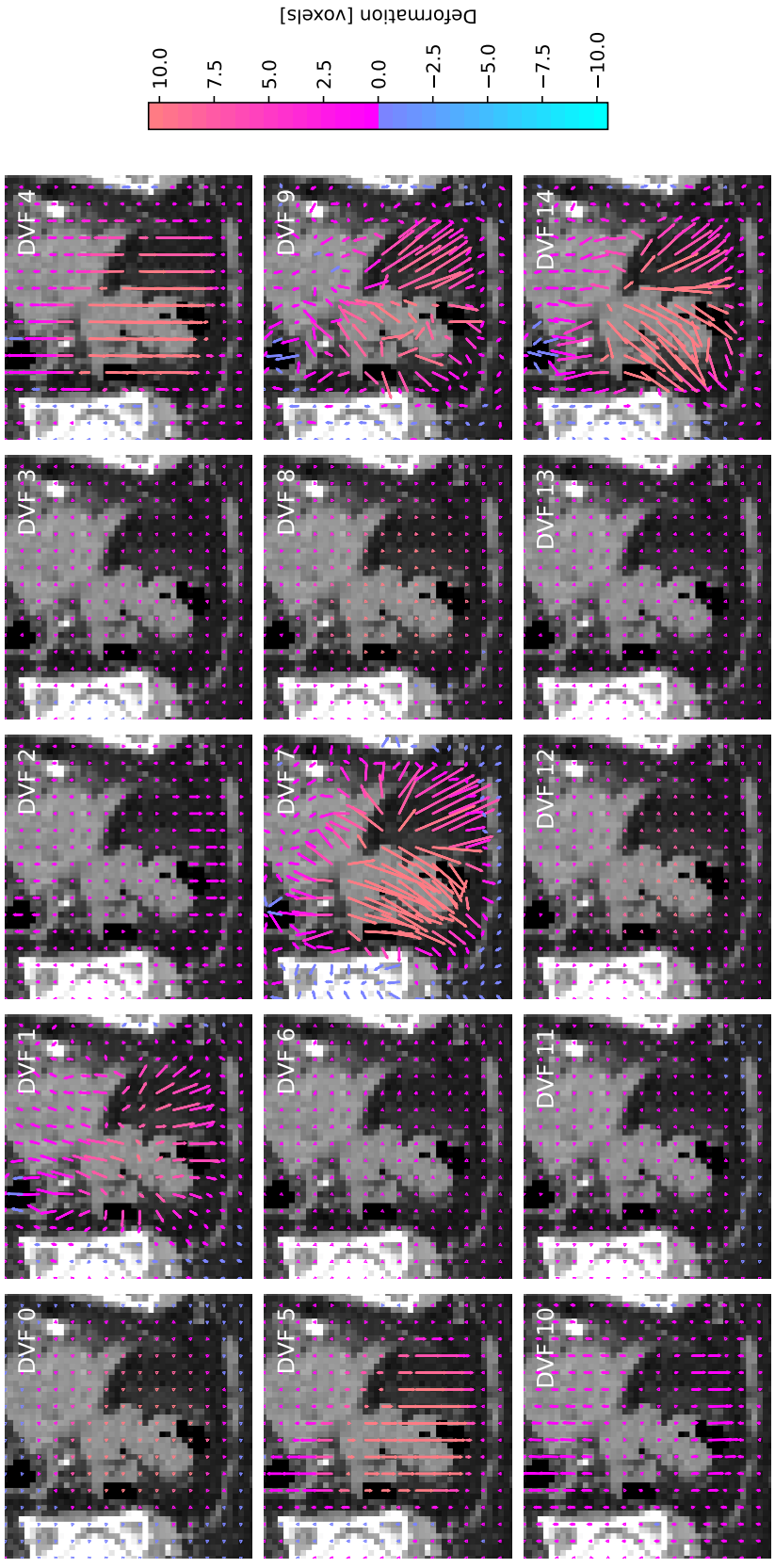



**FIGURE 3.14:** DL-MODIR solutions in the clinical dataset with the target image on the background. Each solution explores a different region of the objective space. Hence, different DVFs are obtained for the same problem.

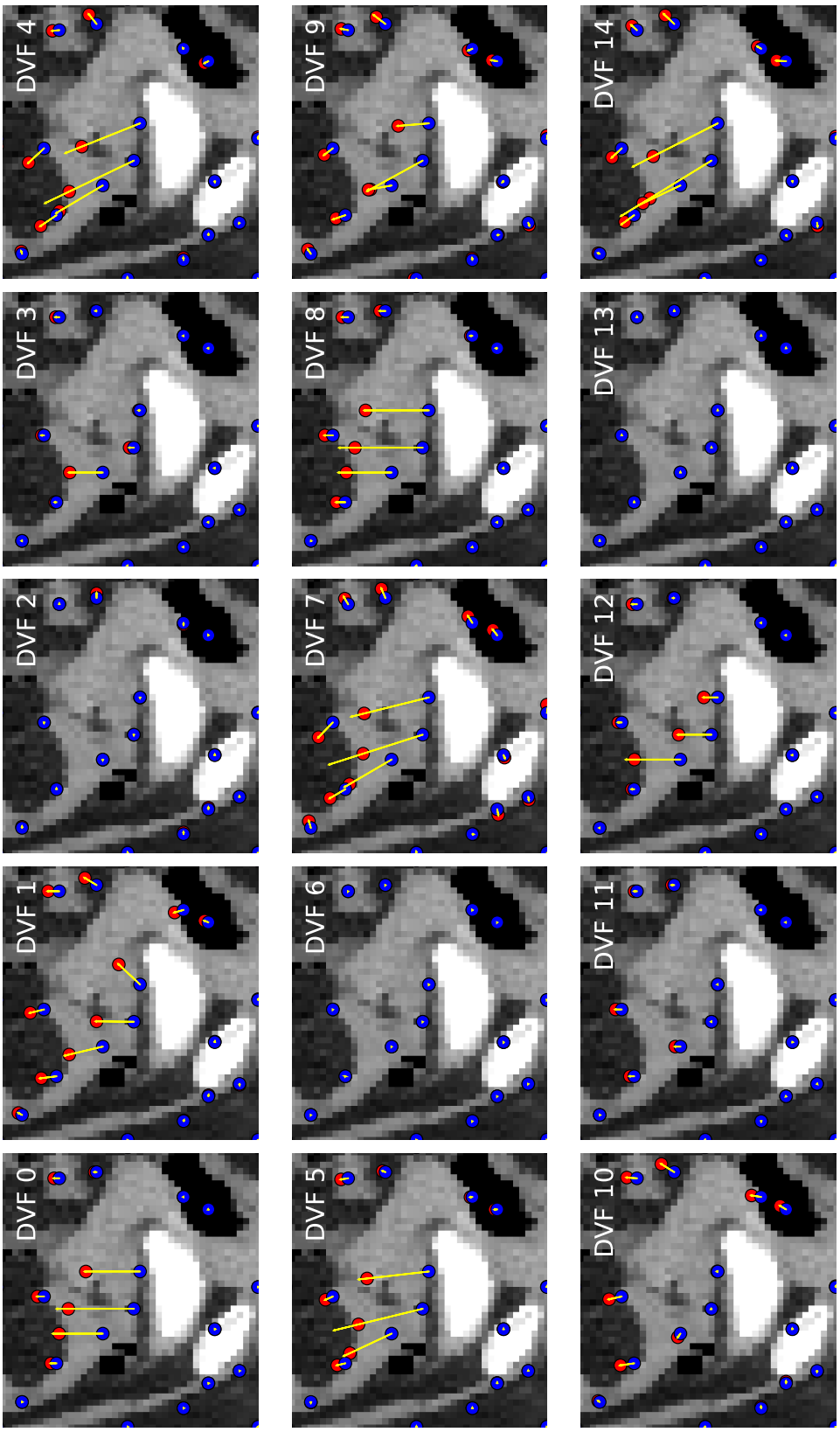



**FIGURE 3.15:** Update of the source mesh grid with the DVFs from DL-MODIR in the clinical dataset. Blue points represent the original mesh grid, and red points represent the mesh grid disrupted with the DVFs. Yellow arrows indicate the displacement vector at the mesh node location, indicating the position the disrupted mesh node should reach. If a disrupted mesh node does not reach the end of the tip, it means that the mesh could not be updated as desired due to constraints.

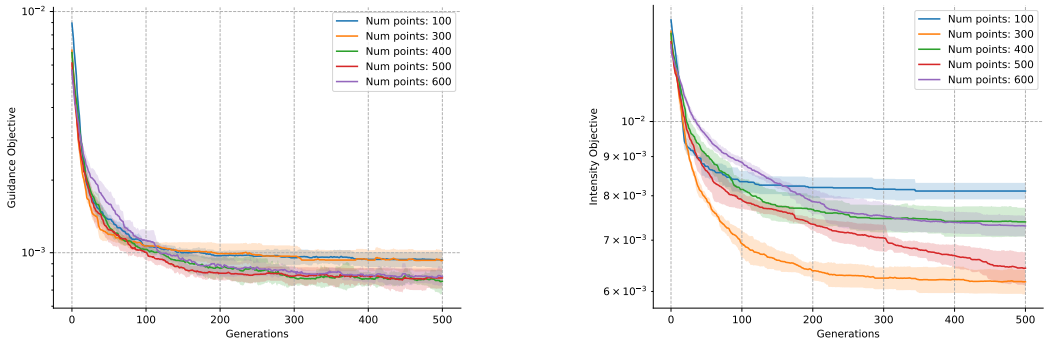
## 3.4 Preliminary experiments

Various preliminary experiments are conducted, the results of which are essential for selecting the best hyperparameters and justifying the implemented strategies in some experiments.

### 3.4.1 Number of nodes

The hybrid method and MOREA are sensitive to hyperparameter tuning. One of the most relevant hyperparameters is the number of mesh nodes. A preliminary experiment to evaluate the optimal value is run in one of the evaluation patients (patient 62). Increasing the number of nodes enhances the granularity of the problem and the susceptibility to constraints caused by folds.

Five different seeds for each evaluated number of nodes (100, 300, 400, 500, and 600) are run in this experiment. To see the effect of this hyperparameter, we simplify the problem by studying the convergence of the exploitation of the guidance and intensity objectives. Figure 3.16 displays the value for the best solution in each objective, comparing the hybrid method with different numbers of mesh nodes.




**(A)** Guidance convergence along generations.

**(B)** Intensity convergence along generations.

**FIGURE 3.16:** Guidance and intensity exploitation along generations for different numbers of nodes in the mesh generation for the developed hybrid method.

A higher number of mesh nodes permits better exploiting the guidance objective, however, no significant difference exists with values higher than 300. Looking at the intensity objective, 300 is the number of nodes that best exploit this objective. Nevertheless, considering both results, the strategy with 500 nodes seems optimal for exploiting both objectives.

### 3.4.2 Percentile selection

The contours of the transformed source image with the different methods are evaluated to analyze the results that the different methods output. We aim to select a percentile of the guidance objective for reproducibility that permits a fair comparison between the different methods. Since the guidance objective is the one that provides the most information on the quality of the solutions, we expect to select solutions with a low guidance objective to visually compare the three methods.

However, overfitting to the guidance (or segmentation) objective can be in detriment of the other two objectives, so we evaluate how different the 90th, 95th, and 98th percentile are from the 100th percentile, referring to the solution with better exploitation of the guidance objective.

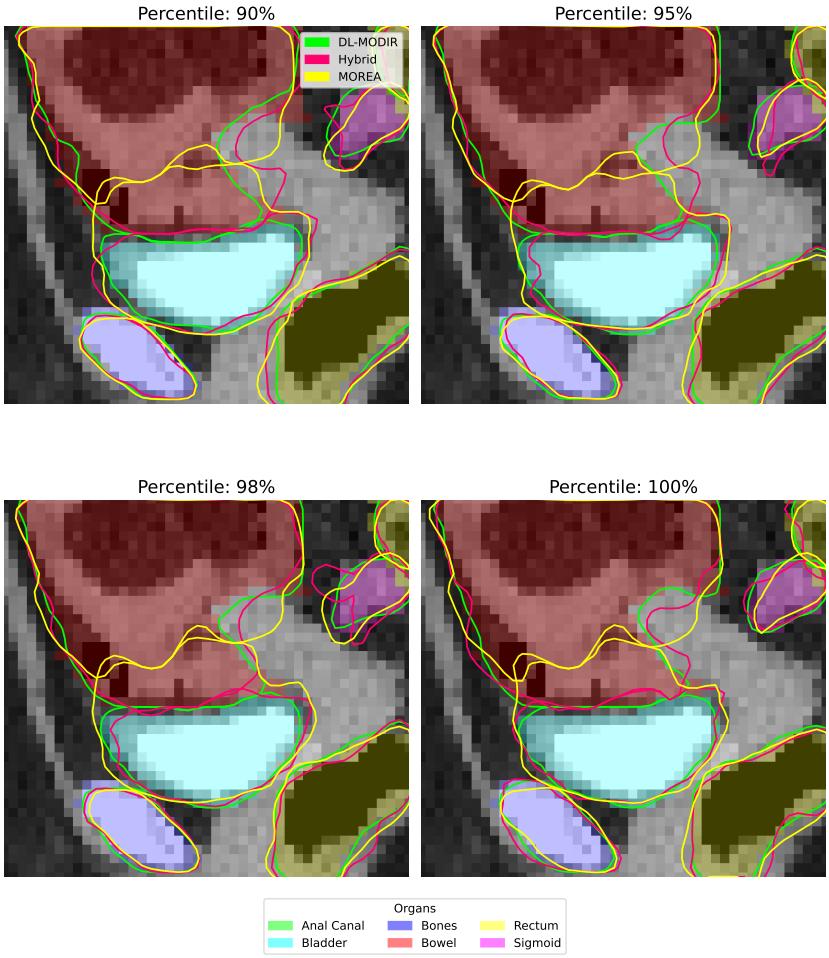
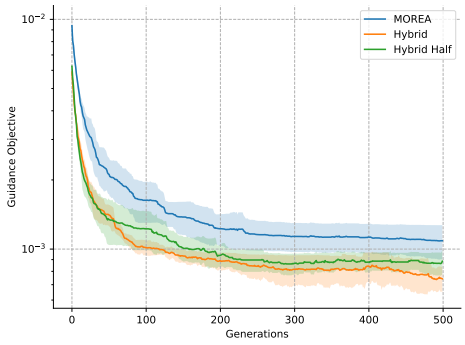



**FIGURE 3.17:** Evaluation of different percentile solutions (90th, 95th, 98th, and 100th) of the guidance (or segmentation) objective in the approximated Pareto front for each method: DL-MODIR, MOREA, and the hybrid method. The three methods show the target image with the segmented organs with the contours of the transformed source masks.

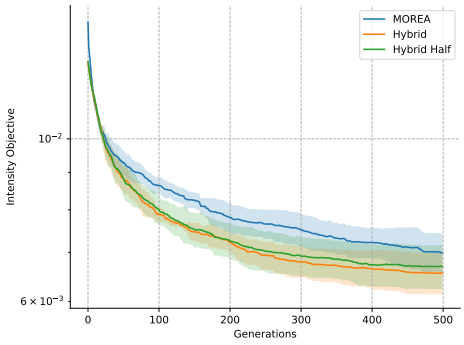
Looking at Figure 3.17, no significant difference is noticeable between the 90th, 95th, 98th, and 100th percentiles. The sigmoid suffers the most significant difference in the segmentation contours, however, the 98th percentile seems worse in this slice and plane, but it is compensated in other organs and slices. However, as there is no significant visual difference between the four percentiles, the 90th percentile is a good choice for comparing the three methods, providing reproducibility, and not overfitting to the guidance objective only. Otherwise, navigating along the approximated Pareto front would be necessary to find the optimal solution on this patient, and the comparison would be subjective depending on the choice.

### 3.4.3 Half-hybrid method

The half-hybrid method proposed in Section 3.2.3 is an alternative to the original hybrid method. This strategy is evaluated in one of the patients (patient 63) by looking at the convergence plots of the exploitation of two objectives, guidance and intensity. Figure 3.18 shows the results of this method compared to MOREA and the hybrid method. The half-hybrid method does not exploit the guidance and intensity objective better than the hybrid method. Thus, providing some solutions generated with Gaussian attractors does not yield to better results in this patient.




**(A)** Guidance exploitation along the generations.




**(B)** Intensity exploitation along the generations.

**FIGURE 3.18:** Half-hybrid method compared to the hybrid method and MOREA for exploiting the guidance and intensity objectives.

## 3.5 Experimental setup

This section delves into the experiments run in this thesis, the used hyperparameters and hardware conditions.

### 3.5.1 Experiments description

Multi-objective problems are complex to analyze since a set of solutions is given instead of a unique solution. Hence, looking at one of the solutions does not provide a global perspective. For this reason, we start looking at the approximated Pareto fronts of the three methods to understand how the given solutions perform on the three objectives and how the objective space is explored.

After having an understanding of the exploration of the different objectives with the Pareto fronts, we look at the convergence of exploiting the three objectives, i.e., their best found value over the course of optimization. This permits to analyze how each method exploits the objectives and its evolution over time. Additionally, it allows to perform a statistical test of which methods have statistically significant improvements in the exploitation of the objectives.

Finally, we look at one of the solutions to analyze it visually. The Pareto front or the objectives convergence is less relevant in clinical practice. Instead, it is essential to look at how close the transformed masks are to the target masks, and, most importantly, how the deformation is shaped. In this third experiment, we study

both situations. The solution in the 90th percentile in the guidance solution is selected according to the preliminary experiment in Section 3.4.2. Three patients are analyzed, two with a significant improvement in the exploitation of the guidance objective and one with no significant improvement, allowing to find a pattern of where the hybrid method performs better than MOREA and, on the other hand, in which cases there is no improvement.

### 3.5.2 Used hyperparameters

The hyperparameters of the three methods are given for reproducing the different experiments. DL-MODIR is run with the following hyperparameters:

| Hyperparameter | Value | Hyperparameter | Value |
|---|---|---|---|
| In-plane size (height × width) | 96 × 96 | Number of segmentation classes | 8 |
| Maximum depth $d_{max}$ | 48 | Number of solutions | 15 |
| Reference point | [1, 1, 1] | Batch size | 1 |
| Number of training iterations | 150,000 | Learning rate | 0.0001 |

**TABLE 3.1:** Hyperparameters used in DL-MODIR.

The hybrid method and MOREA are run with the following hyperparameters:

| Hyperparameter | Value | Hyperparameter | Value |
|---|---|---|---|
| Number of init. steps $N_s$ | 80 | Number of generations | 500 |
| Population size $S_p$ | 600 | Number of clusters | 10 |
| Archive size | 1,000 | Initialization noise factor | 1.0 |
| Number of mesh nodes | 500 | Nodes placement strategy | Surface |
| Organ surface of the mesh | Bladder | Ratio of nodes on surface $r_s$ | 0.25 |

**TABLE 3.2:** Hyperparameters used in MOREA and the hybrid method.

The nodes placement strategy indicates that the mesh nodes are placed on the surface of the indicated organ, which in this case is the bladder. The ratio of nodes on the surface $r_s$ defines the percentage of the total number of mesh nodes placed on the surface of the bladder. This value is set at a 25%. Additionally, MOREA is adapted to place the nodes on the surface of the target image instead of the source image, showing a better performance since an inverse DVF is applied, and the bladder is shrinking. Hence, mapping from the target image is more informative than from the source image.

### 3.5.3 Hardware requirements

Experiments are run on the SHARK HPC cluster at the Leiden University Medical Center. Each experiment is run on one CPU, and one GPU, and requires 30 GB of memory for the run. The experiments used two different GPUs, an NVIDIA GPU RTX A5000 with 24 GB of V-RAM, and an NVIDIA GPU RTX A6000 with 48 GB of V-RAM.

# Chapter 4

# Results

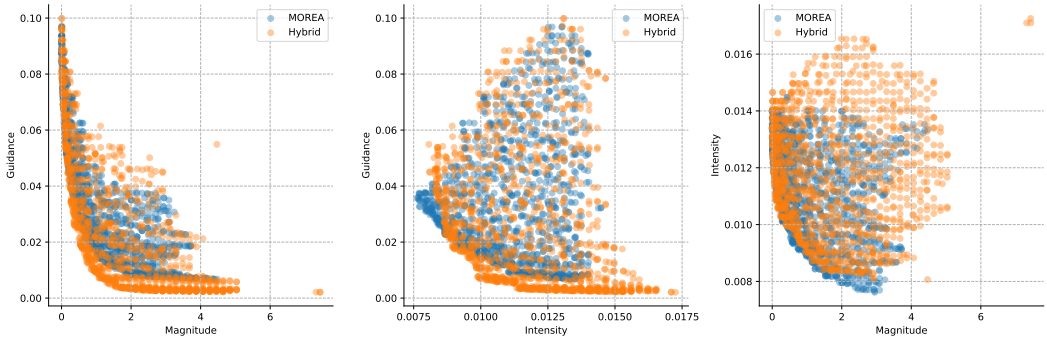
This chapter evaluates the performance of the developed hybrid method to answer the research questions. We begin by assessing the Pareto fronts of the three different methods: DL-MODIR, MOREA, and the hybrid method. We then delve into the convergence of the hybrid method and MOREA for each objective. Finally, we provide a qualitative comparison of the three methods, examining the transformed contours and computed DVFs.

## 4.1 Analysis of the approximated Pareto front

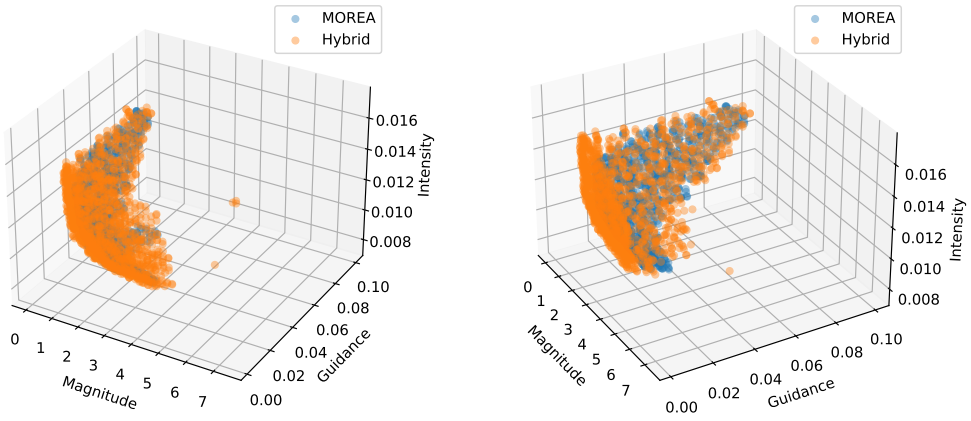
As a multi-objective problem, our methods produce a set of solutions that form the approximated Pareto front. Comparing the approximated Pareto fronts of the different methods is a fair way to understand how the set of solutions of one method compares with another. We compare the hybrid method with MOREA, as both use the same registration quality metrics. However, DL-MODIR uses a different set of metrics, making a direct comparison with the other methods unfair. To guarantee a fair comparison, it would be necessary to evaluate both methods using independent registration quality metrics.

Patient 75 is chosen since it corresponds to the patient in the 50th percentile for the evaluated patients in terms of statistically significant improvement for the hybrid method compared to MOREA in the exploitation of the guidance objective. Figure 4.1 compares the approximated Pareto fronts of the hybrid method and MOREA in two and three dimensions. In this patient, the approximated Pareto front of the hybrid method exploits better the guidance objective since lower values for that objective are achieved. However, there is less exploitation of the intensity objective, whereas MOREA has solutions with lower intensity. For the magnitude objective, the hybrid method has some solutions with a more extensive deformation corresponding to a lower guidance.

For a quantitative comparison, the hypervolume of the approximated Pareto fronts of both methods –hybrid and MOREA– is computed and compared. Table 4.1 shows the $p$-values of a Mann-Whitney U-Test between the hybrid method and MOREA for five runs. The mean increase or decrease in the hypervolume of the hybrid method compared to MOREA is computed with a fixed reference point set at $[10, 10, 10]$ to provide some margin for the higher values. For this patient (patient 75), the hybrid method has a lower average hypervolume than MOREA but without statistical significance. From the 15 evaluated patients, the hybrid method has a statistically significant increase of the hypervolume in 5 cases and a statistically significant decrease in 8 patients. In general, for the patients that have a statistically




**(A)** 2D representation of the approximated Pareto fronts in the three planes of the objective space.




**(B)** 3D representation of the approximated Pareto fronts.

**FIGURE 4.1:** Approximated Pareto fronts comparison for the hybrid method and MOREA in 2D and 3D for patient 75.

significant increase, the percentage of increase is higher than the ones that have a statistically significant decrease. However, since it depends on the patient, it cannot be stated that the hybrid method produces approximated Pareto fronts with higher hypervolumes or the opposite. Nonetheless, it is essential to note that this metric is not relevant in clinical practice, where most clinicians do not understand this value and care more about how the solutions look.

Figure 4.2 shows the approximated Pareto front of DL-MODIR in the same patient but in another objective space. These solutions are used to initialize the hybrid method and intrinsically affect the approximated Pareto front of the hybrid method. Solutions from DL-MODIR are less uniformly spread over the objective space, with most solutions concentrated around the same region. The solutions seem clustered in the corner with high intensity and segmentation, and low magnitude, where small or no deformation occurs. This is the same problem encountered in the original DL-MODIR paper [25].

| Patient | $p$-value | % vs MOREA |
|---|---|---|
| Patient 62 | **0.008** | **↓ 0.026%** |
| Patient 63 | **0.008** | **↓ 0.015%** |
| Patient 64 | **0.016** | **↑ 0.096%** |
| Patient 65 | **0.008** | **↑ 0.191%** |
| Patient 66 | **0.008** | **↑ 0.117%** |
| Patient 67 | **0.008** | **↓ 0.013%** |
| Patient 68 | **0.008** | **↓ 0.023%** |
| Patient 69 | **0.008** | **↓ 0.019%** |
| Patient 70 | **0.008** | **↓ 0.034%** |
| Patient 71 | **0.008** | **↓ 0.017%** |
| Patient 72 | **0.008** | **↑ 0.552%** |
| Patient 73 | 0.095 | ↓ 0.037% |
| Patient 74 | **0.032** | **↑ 0.053%** |
| Patient 75 | 0.421 | ↓ 0.007% |
| Patient 76 | **0.008** | **↓ 0.026%** |

**TABLE 4.1:** Hypervolume comparison for five different seeds. $p$-values are computed from a Mann-Whitney U-Test between the hybrid method and MOREA. The percentage compares the increase (green) or decrease (red) of the mean of the hybrid method distribution with the MOREA distribution. Bold values represent a statistically significant difference ($p$-value $< 0.05$).

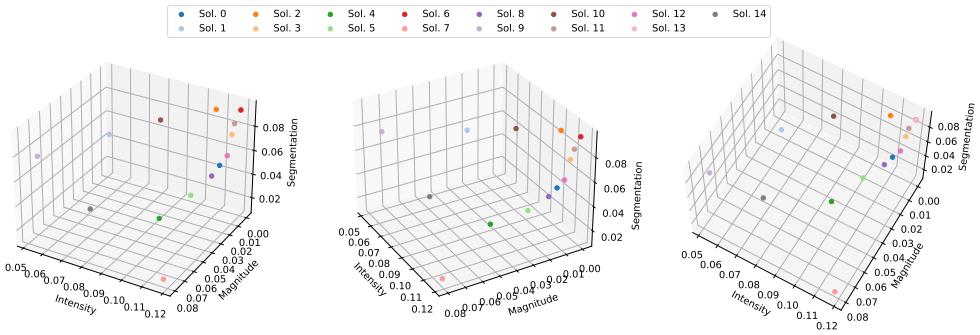
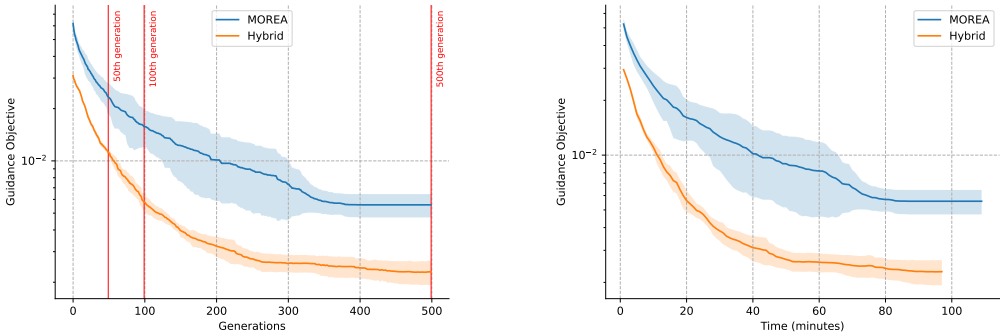



**FIGURE 4.2:** Approximated Pareto front of DL-MODIR for patient 75.

## 4.2 Comparison of the objectives convergence

After visualizing the entire set of solutions, it is essential to study how each method exploits each objective and converges. The solution of the approximated Pareto front with the lowest value for each objective –exploitation of the objective– is used for the convergence comparison. However, this is just a partial visualization of the solutions provided by the different methods and does not outline how good they are compared to others. The qualitative comparison at the end of this chapter permits a better understanding of how both methods work by visually inspecting their solutions.

Despite working in a multi-objective problem, our convergence comparison focuses on the guidance objective since it gives the best qualitative indication of how close the transformed source image approximates the target image. The intensity registration quality metric used –the mean squared difference– is not robust enough for uncalibrated images, for instance, when a contrast agent is used, as is the case in some of the evaluated patients. The third objective, the magnitude, only represents the solution with the lowest deformation in the population, which is not informative enough since we are interested in large deformation as we know that the bladder compresses from the source image to the target image.




**(A)** Guidance convergence along generations.

**(B)** Guidance convergence along time.

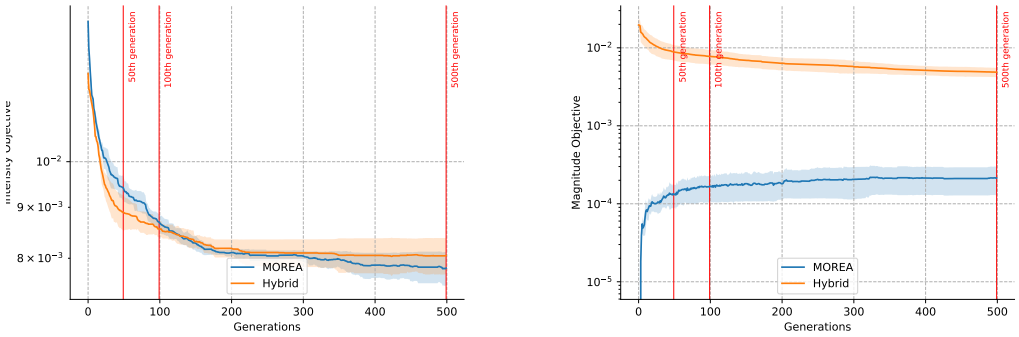
**FIGURE 4.3:** Guidance convergence for the best solution in this objective along generations and time for patient 75.

Five random seeds are run for each method and patient to overcome stochasticity. The same patient as the previous experiments (patient 75) is used to compare both methods. However, results on the rest of the patients exploiting the guidance objective can be found in the Appendix (Figure B.1). Figure 4.3 displays the guidance evolution along the number of generations and time. The guidance objective of the hybrid method at the start is lower than that of MOREA. The initial population, generated from DVFs given by DL-MODIR has a better exploitation of the guidance than MOREA and stays lower throughout the 500 generations.

Additionally, the hybrid method is more stable in its results –lower standard deviation over the runs–, giving less room to stochasticity for exploitation of the guidance objective. Ultimately, both methods achieve a plateau, meaning that convergence has been achieved and cannot better exploit the guidance objective. The guidance over time shows that the hybrid method is faster than MOREA, and even if the program is terminated early at any moment, the hybrid method will outperform MOREA in every case for this patient.

Figure 4.4a shows the convergence for exploiting the intensity objective, where both methods converge similarly for this objective. On the other hand, the magnitude objective exploitation in the hybrid method is always higher than MOREA. This means that the solution with lower deformation of the hybrid method is always higher than the one in MOREA, which is not informative to evaluate the quality of a set of solutions.

Table 4.2 shows the $p$-values of a Mann-Whitney U-Test on the best solution for the guidance objective in the hybrid method and MOREA at three timesteps: 50,




**(A)** Intensity convergence

**(B)** Magnitude convergence

**FIGURE 4.4:** Intensity and magnitude convergence for the best solution of each objective along generations for patient 75.

100, and 500 generations, the latter representing the end of an entire optimization. This significance test does not assume normality in the distribution, being more robust when the distribution of results is unknown, as it is the case. The hybrid method shows a statistically significant decrease in the mean guidance objective in 10 of the 15 patients after 50 iterations and 8 of the 15 validation patients at the end of the optimization. Most patients have a significant improvement at the beginning, leveraging information from the DL-MODIR DVFs. Nevertheless, in one patient (patient 68), the statistical significance difference at the beginning is later counterbalanced by MOREA at the end of the run (500 generations). These results demonstrate that the hybrid method has a better or comparable exploitation of the guidance objective at the beginning and the end of the optimization, with any patient having a statistically significant increase at any evaluated generations.

| | 50 generations | | 100 generations | | 500 generations | |
|---|---|---|---|---|---|---|
| **Patient** | ***p*-value** | **% vs MOREA** | ***p*-value** | **% vs MOREA** | ***p*-value** | **% vs MOREA** |
| Patient 62 | **0.016** | ↓ **33.4%** | **0.008** | ↓ **37.8%** | **0.016** | ↓ **31.7%** |
| Patient 63 | 0.222 | ↑ 7.8% | 0.421 | ↑ 2.8% | 0.548 | ↑ 6.1% |
| Patient 64 | **0.008** | ↓ **43.9%** | **0.008** | ↓ **52.6%** | **0.008** | ↓ **67.0%** |
| Patient 65 | **0.008** | ↓ **37.3%** | **0.008** | ↓ **50.4%** | **0.008** | ↓ **45.5%** |
| Patient 66 | **0.008** | ↓ **53.5%** | **0.008** | ↓ **66.1%** | **0.008** | ↓ **76.6%** |
| Patient 67 | 0.151 | ↓ 6.5% | 0.310 | ↓ 5.4% | 0.056 | ↓ 9.9% |
| Patient 68 | **0.008** | ↓ **31.6%** | **0.008** | ↓ **38.4%** | 0.222 | ↓ 8.8% |
| Patient 69 | 0.151 | ↓ 16.6% | 0.095 | ↓ 18.1% | 0.151 | ↓ 12.8% |
| Patient 70 | 0.056 | ↓ 11.3% | 0.095 | ↓ 7.3% | 1.000 | ↓ 0.5% |
| Patient 71 | 0.690 | ↓ 2.5% | **0.016** | ↓ **7.3%** | 0.151 | ↓ 5.8% |
| Patient 72 | **0.008** | ↓ **47.4%** | **0.008** | ↓ **53.8%** | **0.008** | ↓ **70.2%** |
| Patient 73 | 0.095 | ↓ 23.6% | 0.222 | ↓ 13.5% | 0.690 | ↑ 1.6% |
| Patient 74 | **0.008** | ↓ **41.8%** | **0.008** | ↓ **53.4%** | **0.008** | ↓ **55.0%** |
| Patient 75 | **0.008** | ↓ **51.9%** | **0.008** | ↓ **63.3%** | **0.008** | ↓ **59.0%** |
| Patient 76 | **0.016** | ↓ **21.3%** | **0.008** | ↓ **28.5%** | **0.008** | ↓ **24.7%** |

**TABLE 4.2:** *p*-values of a Mann-Whitney U-Test between the hybrid method and MOREA, and the percentage comparison of the mean of the hybrid method vs MOREA. Both metrics are evaluated in the guidance objective after 50, 100, and 500 generations. Green means a decrease in the hybrid mean compared to MOREA and red means an increase, being bold when the increase or decrease is statistically significant.

## 4.3 Qualitative assessment of methods

Numbers provide only a partial understanding of the results and do not directly translate to clinical practice. The main goal of this thesis is to provide a method that works in real-world scenarios. Therefore, a qualitative comparison permits going beyond and visualizing each method's provided solutions. According to Section 3.4.2, we compare the 90th percentile solution in the guidance objective front for each method: hybrid, MOREA, and DL-MODIR.

To better comprehend the problem, we visualize three patients, two with statistically significant improvement exploiting guidance after 500 generations and one with no statistically significant difference.

### 4.3.1 Study of median performance: a case study of patient 75

Patient 75 corresponds to the 50th percentile of the patients with statistically significant improvement exploiting the guidance objective (Table 4.2), representing an average case where the hybrid method outperforms MOREA in guidance exploitation.

Figure 4.5 shows how the three methods compare regarding guidance for the solution in the 90th percentile. The DL-MODIR contour matches with the segmentation mask, while the hybrid method and MOREA struggle when this deformation is considerable, such as in the bowel and the bladder. However, the hybrid method can better capture this large deformation by getting closer to the final size of the bladder. This large deformation incapability of MOREA can be due to the impossibility of the mesh points moving further down due to the constraints this model works with to avoid folds. Moving a region of MOREA in the same direction can be

stopped due to folding constraints, especially when mesh nodes are not moving altogether in the same direction. In the hybrid method, however, this is solved when we apply the DVF at the initialization, and the whole organ moves down simultaneously in a multiple-step process to overcome constraints, permitting it to capture this large deformation and, therefore, have a closer guidance. Other static organs, such as the bones, are better registered since no deformation occurs between the source and target image.

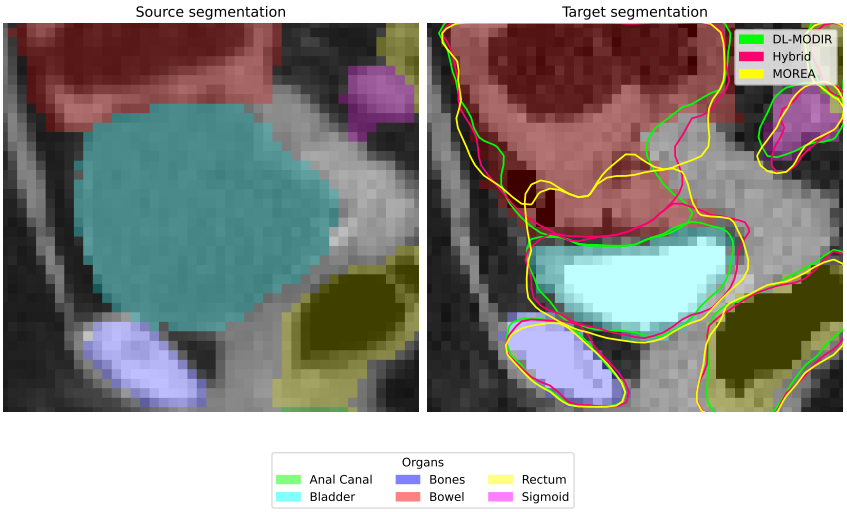



**FIGURE 4.5:** Sagittal plane of patient 75, showing the source and target images with segmentation masks for each organ as provided by the radiologist. The target image includes transformed organ contours from the three methods for the solution in the 90th percentile of the guidance objective.

Another important fact is the cropping of the image during pre-processing. Thus, some parts of the organ in the target image are out of the field of view in the source image. For instance, the bowel occupies a larger space in the target image, where some appearing parts are cut from the source image. It is essential to remark that the target segmentation masks cannot be considered ground truth since results may change depending on the radiologist [65], and overfitting them may lead to wrong transformations.

Closer guidance does not necessarily always mean a better registration. We can reach the same guidance results by performing different deformations. In clinical practice, therefore, it is more important to know how this deformation takes place than to focus solely on the transformed segmentation masks. Figure 4.6 shows the DVFs for the three methods on the same patient and slices. The resulting DVF in the hybrid method has longer arrows that yield larger deformations. As an inverse DVF, the arrow's tip points to the corresponding voxel in the source image for the pixel located at the arrow's origin. Since the bowel compresses the bladder down, points in the bladder and the bowel are pointing up, corresponding these voxels to the upper voxels in the source image. The third dimension, perpendicular to the plane, is expressed in colors.

The DL-MODIR DVF seems less natural as it contains drastic direction changes in neighbor arrows. This occurs in the right of the bowel, where voxels in the bowel point to the image's top left and regions on the right of the bowel point to the top right.

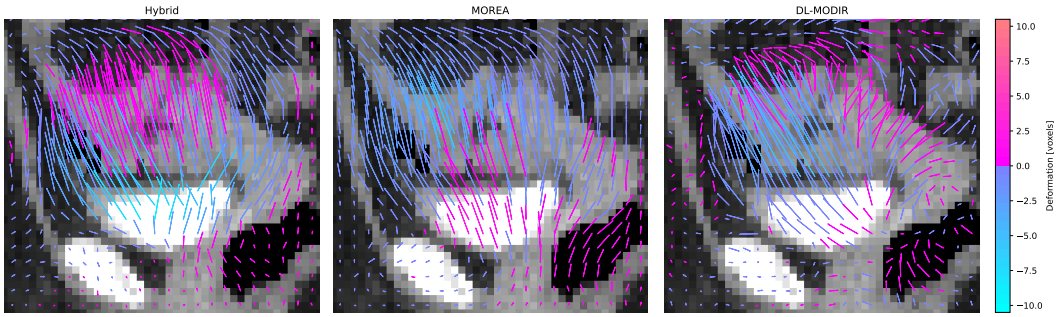



**FIGURE 4.6:** Sagittal plane of patient 75, displaying the inverse DVFs for the three methods: DL-MODIR, MOREA, and the hybrid method. Arrow colors indicate the direction of the DVF in the third dimension.

Focusing on another anatomical plane, the coronal (Figure 4.7), the guidance behaves similarly. In this case, however, the region of the bowel out of the field of view is even more significant, so expanding a non-visible region is even more complicated.

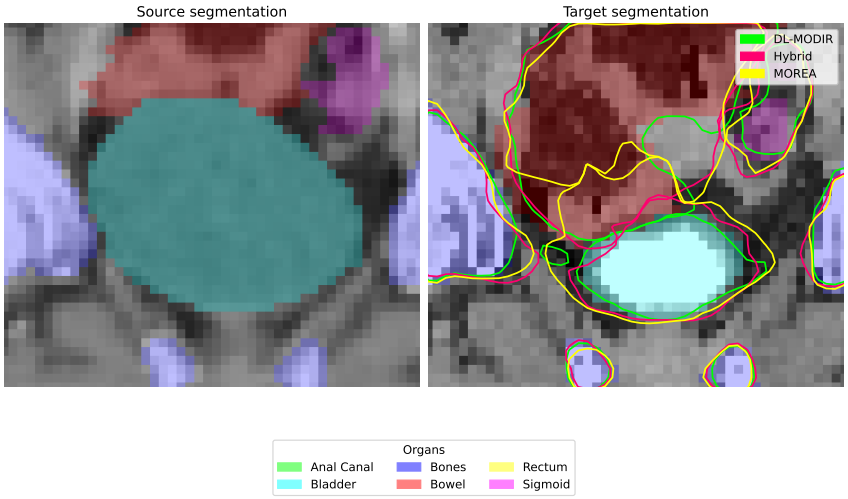



**FIGURE 4.7:** Coronal plane of patient 75, showing the source and target images with segmentation masks for each organ as provided by the radiologist. The target image includes transformed segmentation contours from the three methods for the solution in the 90th percentile of the guidance objective.

More conclusions can be extracted from the image by looking at the DVFs (Figure 4.8). The deformation of the bones, however, is better captured by DL-MODIR. No deformation occurs in DL-MODIR since the model computes independently a

displacement vector for each voxel. Another reason is that DL-MODIR gives more importance to the intensity metric than MOREA and the hybrid method since it optimizes by maximizing the hypervolume of the approximated Pareto front. On the other hand, the hybrid method and MOREA show some deformation in bone structures. This can be due to multiple reasons: the granularity of the problem means that, depending on the position of the mesh nodes, some deformation occurs in bone structures due to a deformation of the neighbor structures. In Figure 4.7, the bladder is very close to the bone, and a movement of a node in the bladder is translated into a movement of the bone. The other reason is that MOREA and the hybrid method differ from the original MOREA paper [4] in considering the biomechanical properties of the organs. Compared to the original MOREA, the current method does not know the biomechanical properties of the organs, and deforming a bone structure is as difficult as deforming the bladder, even though this is not the case in reality.

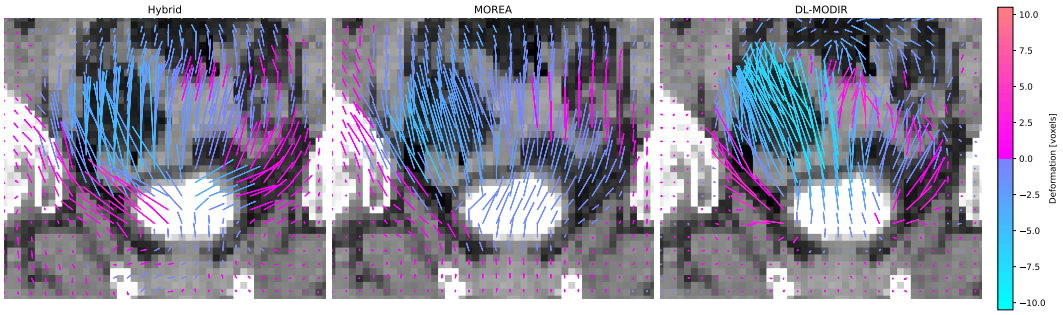



**FIGURE 4.8:** Coronal plane of patient 75, displaying the inverse DVFs for the three methods: DL-MODIR, MOREA, and the hybrid method. Arrow colors indicate the direction of the DVF in the third dimension.

### 4.3.2 Optimal guidance exploitation: a case study of patient 66

To better understand when the hybrid method has a better exploitation of the guidance objective, we visualize the results on the patient where the mean guidance objective in the hybrid method decreased the most compared to MOREA after 500 generations, corresponding to patient 66, with a 76.6% decrease (Table 4.2). Figure 4.9 shows that the most significant difference occurs when the deformation is larger. While the hybrid method can reach a bladder contour similar to the target segmentation mask, MOREA cannot deform completely. Thus, the difference in guidance exploitation seems higher for the patients where the bladder compression is more extensive. In this patient, the same problem occurs with the field of view, where most of the bowel present in the target segmentation does not appear in the source segmentation.

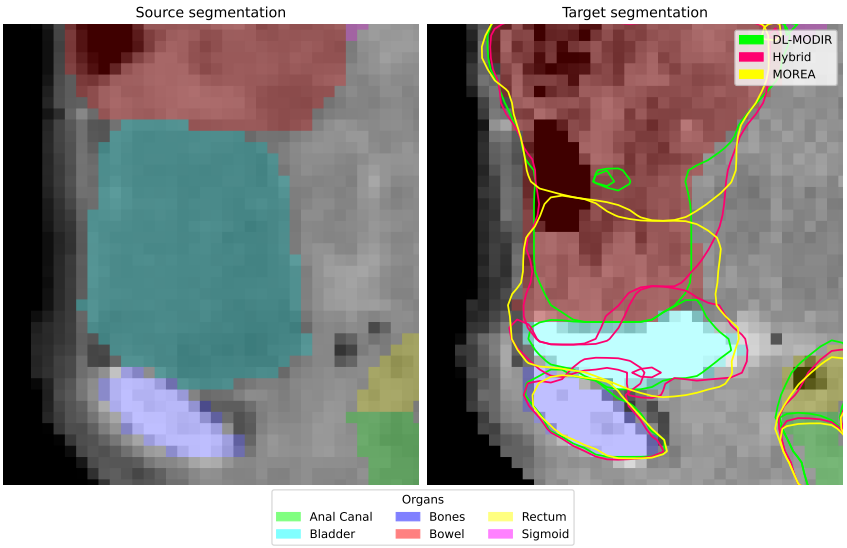



**FIGURE 4.9:** Sagittal plane of patient 66, showing the source and target images with segmentation masks for each organ as provided by the radiologist. The target image includes transformed segmentation contours from the three methods for the solution in the 90th percentile of the guidance objective.

DVF results are similar to the previous patient. However, in this example, DL-MODIR shows a bigger deformation of the top of the bladder, but the lower part has minimal deformation. Hence, compression only occurs on top while the rest of the bladder remains static. This is seems less anatomically realistic since we expect a more distributed compression across the entire organ. On the other hand, MOREA shows a deformation more distributed across the entire bladder but cannot reach the large deformation between the source and the target image. The hybrid method, since it uses solutions from DL-MODIR, shows a similar behavior to DL-MODIR, with a larger deformation in the top of the bladder and less uniformly spread across the entire organ.

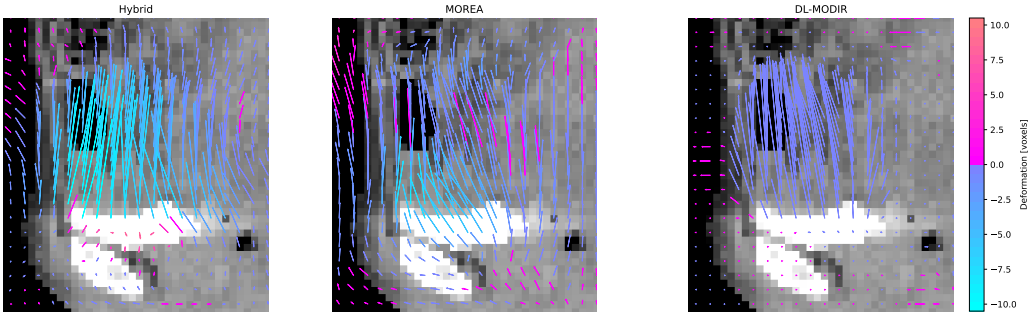



**FIGURE 4.10:** Sagittal plane of patient 66, displaying the inverse DVFs for the three methods: DL-MODIR, MOREA, and the hybrid method. Arrow colors indicate the direction of the DVF in the third dimension.

### 4.3.3 Analysis of non-significant improvement: a case study of patient 63

Finally, a patient with no statistically significant improvement is evaluated, patient 63, to understand in which cases the hybrid method performs similarly to MOREA. Figure 4.11 shows the transformed source contours of each method in the sagittal plane for this patient. In this case, DL-MODIR contours continue matching perfectly the segmentation masks, while the other two methods overfit less on the guidance objective. Focusing on the bladder, MOREA seems to better match in some regions of the bladder, such as the bottom-right lobe, while the hybrid method captures better the left lobe. However, in general terms, no significant difference is visible in their transformed segmentation mask, meaning that both methods perform similarly when the bladder deformation is smaller or nonexistent.

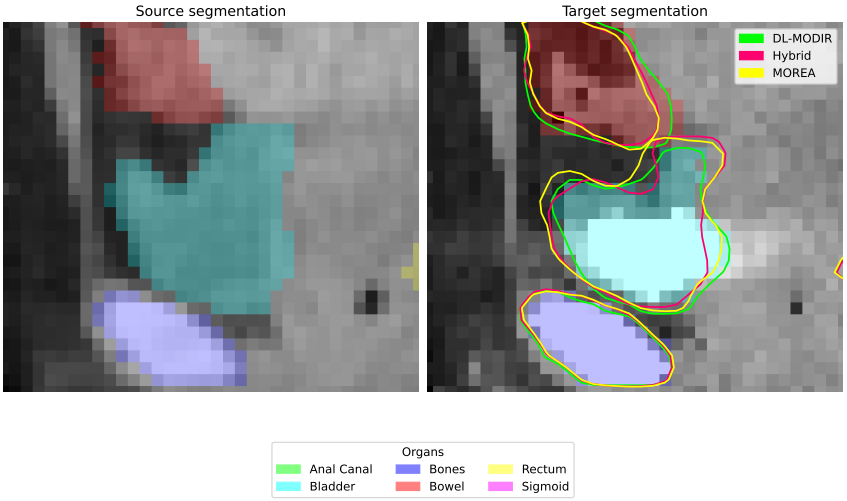



**FIGURE 4.11:** Sagittal plane of patient 63, showing the source and target images with segmentation masks for each organ as provided by the radiologist. The target image includes transformed segmentation contours from the three methods for the solution in the 90th percentile of the guidance objective.

Focusing on the DVFs (Figure 4.12), the hybrid method and MOREA look similar, meaning that DL-MODIR has a similar solution to the one approached by MOREA or that solutions with lower deformations, closer to the ones that MOREA uses, survive in later generations.

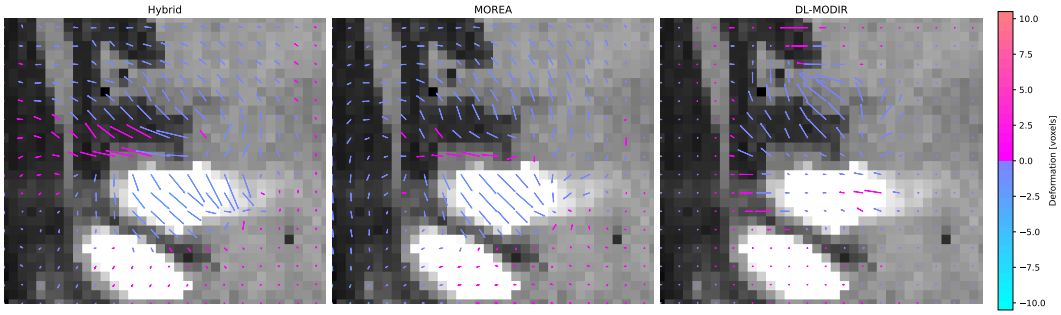



**FIGURE 4.12:** Sagittal plane of patient 63, displaying the inverse DVFs for the three methods: DL-MODIR, MOREA, and the hybrid method. Arrow colors indicate the direction of the DVF in the third dimension.

## 4.4 Grouping patients

We can group patients into two groups by attending to the evaluated solutions and looking at each patient solution individually in Appendix C. One group for patients with large bladder deformation –the difference in the bladder size between the source image and target image is visually significant– and a second group when the difference is slight.

The large bladder deformation group includes patients 64, 65, 66, 72, 74, and 75. After 500 generations, all these patients show a statistically significant improvement in guidance exploitation (Table 4.2), meaning that the hybrid method is better in all cases where a large deformation occurs in the evaluated patients.

Conversely, the hybrid method's effectiveness is less pronounced in the second group of patients (62, 63, 67, 68, 69, 70, 71, 73, and 76). Only two patients, 62 and 76, show a statistically significant improvement in the guidance objective for the hybrid method, however, they show the lowest mean decrease compared to MOREA. The remaining patients do not show a statistically significant improvement, suggesting that the hybrid method may be less effective when the bladder deformation is small. In these cases, MOREA can achieve comparable results.

# Chapter 5

# Discussion

This chapter discusses the reported results by answering to the research questions, highlights the limitations of the developed hybrid method, and proposes various avenues for future work to improve hybrid methods.

## 5.1 Approximated Pareto fronts comparison

For a quantitative comparison of the introduced hybrid method with MOREA, the approximated Pareto fronts are analyzed. The experiments demonstrate that the set of solutions from the hybrid method is uniformly spread across the entire objective space, similar to MOREA. In contrast, DL-MODIR explores the objective space to a lesser extent, with most solutions concentrated in the region of small deformation, as reported in the original paper [25]. Despite the non-uniform spread nature of DL-MODIR solutions, their use in the hybrid does not affect the distribution of the approximated Pareto front of the hybrid method, which remains comparable to MOREA. When evaluating the hypervolume, neither MOREA nor the hybrid method consistently dominates the other, with results varying depending on the patient.

DL-MODIR cannot be directly quantitatively compared with the other two methods as they use different registration quality metrics to evaluate the objectives, leaving an open door for future research.

## 5.2 Convergence comparison

Convergence plots focus on exploiting the different objectives, i.e., analyzing the lowest solution in each objective and its evolution along generations. As a data series, temporary features can be extracted from it. For instance, Appendix B shows the convergence of the guidance objective for all patients. After 500 generations, all evaluated patients show convergence, reaching a plateau in exploiting the three objectives. The application of DL-MODIR solutions in the hybrid method gives MOREA an advantage by starting with an initial population that outperforms the Gaussian attractors initialization. This highlights the impact of the warm start on the performance of the initial population. But, most importantly, the warm start permits exploiting the guidance to a bigger extent than MOREA, where the Gaussian attractors initialization falls into a local optimum for the guidance objective. Hence, the hybrid method is comparable to or better than MOREA exploiting the guidance objective in the evaluated patients. All patients with large deformation showed a statistically significant improvement in the guidance exploitation. On

the other hand, the hybrid method only has statistically significant improvement in two of the nine evaluated patients with small deformation, with comparable results in the rest.

In terms of intensity, the hybrid method does not outperform MOREA. However, this does not imply that the solutions are inferior. The intensity objective is not sufficiently robust, as it relies on the mean squared difference for the evaluation. DL-MODIR, for instance, is more robust in measuring this objective by using the normalized cross-correlation, which can deal with uncalibrated images. This, added to the fact that target images of the evaluated patients contain a contrast agent, which produces a brighter bladder intensity compared to the source image, results in a significant intensity difference between both images. Thus, they are labeled as poor solutions in the intensity objective despite correctly matching pixels from the transformed source image into the target image. Evaluating MOREA and the hybrid method on another intensity registration quality metric, such as the normalized cross-correlation, could improve the results, offering a promising avenue for future research.

While evaluating the solution with the lowest magnitude, the hybrid method always starts with a higher value than MOREA. The study of the magnitude is not needed as it is a trivial task. For instance, a solution that exploits the guidance objective is a solution with the original mesh configuration. Therefore, no findings can be deduced from this result apart from the fact that the lowest deformation applied with the DVFs is always more aggressive than the lowest deformation with Gaussian attractors.

The study of the convergence plots permits us to answer our research question related to convergence acceleration. Looking at the results we cannot state that the introduced hybrid method accelerates convergence since both methods reach a plateau at a similar number of generations. However, we can say that the introduced hybrid method exploits better the guidance objective in most patients, leading to solutions with a smaller distance between the transformed source contour and the target contour.

## 5.3 Visual interpretation limitations

To answer the last research question, related to a qualitative comparison, we compared the transformed source contours and the DVFs of the three methods. The 90th percentile of exploiting the guidance objective is selected for each studied patient. Ideally, it would be necessary to iterate over the solutions to find the optimal for each patient. In real-world scenarios, no fixed solution works for all patients, and letting the user decide the best option permits tuning the solution based on different contexts. However, this task is complex –as it requires anatomical knowledge– and time-consuming. For these reasons, we consider that choosing a fixed solution of the approximated Pareto front is better for reproducibility, despite solutions very close in the guidance objective can have significant differences in the other two objectives, thus resulting in very different deformations.

It is important to note that in the goal of cervical cancer RT planning, where the delivered dose distribution of each organ is aimed to be adapted from the source to the target image, the contours provide partial information of the registration.

However, the deformation of the organ's interior and the inter-organ space is not evaluated with such an objective. On the other hand, the magnitude and intensity objectives can capture it. For instance, the intensity metric looks for the visual similarity of the inter-organ regions, but the fact that images are uncalibrated hinders DIR methods.

The expansions of small regions into more prominent organs in DL-MODIR, or the particular interest of the methods to deform the contours of the organ without paying attention to the interior of the organ, can have a negative impact in clinical practice. This issue is especially challenging when the organ does not contain texture and has a uniform intensity. However, the importance of the DVF in the interior of the organ depends on the situation. In some cases, the delivered dose is uniform across the entire organ, which does not affect the quality of the DVF in the interior. However, when there is a sudden significant difference in the delivered dose on two close regions in the interior of the organ, the DVF can affect the patient's outcome. Additionally, the interior of some organs like the bladder or the rectum can be filled with substances, producing a change in the voxels' intensity and not being necessary to predict the deformation of urine or feces as they are dynamic elements.

In contrast, DL-MODIR over-adjusts its solutions to the contours, as shown in Figure 4.5, where the transformed source contours distance with the target mask is low. Considering that segmentation masks are subjective and depend on the radiologist, solutions that have very good results for the guidance objective do not relate to better deformations. MOREA and the hybrid method do not match as closely to the contours of the organs, but when large deformation occurs, they are not able to achieve full deformation.

Referring to the DVFs, the ones of MOREA and the hybrid method end to follow the modeled object more closely, performing deformations throughout the entire organ rather than just at the contours, as it happens with DL-MODIR. Some solutions in DL-MODIR produce DVFs with a deformation close to zero in one axis, resulting in a uniform distribution with arrows pointing in the same direction (Figure 3.14). When analyzing the reason for this occurrence, we found that DL-MODIR in the original paper showed results with the same behavior. Therefore, only two of the three dimensions deform, while at the same time is able to reach optimal guidance values. Evaluating other papers that used VoxelMorph for DIR, we did not find this behaviour in any of them [38, 29]. Hence, we think this behavior can be due to the multi-objective landscape of DL-MODIR as not all solutions show this pattern, but further research is needed. Another reason may be the anatomical structure analyzed in this thesis and the original DL-MODIR paper, where both methods study the abdominal region with focus on the bladder. Yet, FEM-based models behave more similarly to the human body, moving structures jointly and not independently computing each DVF's displacement vector as DL-MODIR does. This is noticeable by looking at the interrelation between neighbor displacement vectors where in MOREA and the hybrid method we see less sudden changes in neighbor voxels and the deformation seems more smooth than DL-MODIR.

Lastly, the DVF analysis is done from an inexperienced perspective. Understanding the complex movements in the human body and their realism requires anatomical expertise, which could not be achieved in the period this thesis was

developed. Alternatively, experts' opinions in the field would be invaluable in performing a deeper and more reliable analysis of the DVFs. Ideally, feedback from multiple experts would be necessary to deal with the variability of different opinions. At the same time, DVFs can be abstract and difficult to interpret in two dimensions, hindering the comparison of different solutions. Landmarks, for instance, are more interpretable since they connect relevant anatomical points from the source image to the target image, but further research needs to be done in this field to enhance DIR interpretability.

## 5.4 DL-MODIR adaptability

The original DL-MODIR paper was evaluated on a distinct problem with Magnetic Resonance Imaging (MRI) scans instead of CT scans. This posed a challenge since the pre-trained model was not useful for our clinical dataset, which also contains a different input size. DL-MODIR successfully resolved DIR in CT scans as happened in a previous study that already demonstrated the feasibility of VoxelMorph –the neural network DL-MODIR is based on– on abdominopelvic CT scans [64].

Furthermore, the dataset DL-MODIR was trained for in the original paper, contained small deformations contrasting with the large deformations occurring in some of our patients. Despite these differences, DL-MODIR showed adaptability to a different setting, performing large deformations on CT scans and resulting in a promising method to be used in multiple settings.

## 5.5 MOREA differences

The solutions found by MOREA differ from those in the original paper [4], where the transformed contours of the source bladder get closer to the target bladder despite the large deformation occurring. This can be due to multiple reasons. First, the version of MOREA used in this thesis does not consider the biomechanical properties of the organs. Therefore, the bladder deformation in the original paper can be less costly than in the used MOREA, but further analysis is necessary to understand the impact of the results. Furthermore, the developed method does not perform adaptive steering. Adaptive steering helps MOREA perform larger deformations, where only solutions with the best guidance survive after several generations. However, the use of adaptive steering comes with some drawbacks. The threshold of surviving solutions needs to be manually tuned, increasing the hyperparameter space and, thus, the complexity of the method. Additionally, they introduce bias in the selected solutions, forcing large deformations and discarding promising solutions. A more extensive analysis of the effect of adaptive steering and the biomechanical properties of the organs is necessary to understand their impact on MOREA and how the hybrid method compares to MOREA under these conditions.

## 5.6 Hardware limitations

Like its predecessors, the proposed method has some limitations that challenge its application for clinical practice. For online registration, performed while the patient is static, this process needs to be done in real-time. For a better understanding of the applicability of the developed hybrid method in this context, it is necessary to study the optimized solutions after a short optimization with a few generations. Convergence plots suggest that the biggest change is performed at the beginning, with an exponential decay function behavior and minimal changes occurring after 100 iterations in most cases. Additionally, the statistical test performed on the exploitation of the guidance objective indicates that the hybrid method also outperforms MOREA at the first stages, after 50 and 100 generations. Thus, solutions given after a short period can be used for this purpose. However, in cancer planning, offline RT is also possible. In this case, the re-planning is done hours or days before the patient gets the treatment and real-time DIR is not necessary anymore, allowing systems to perform a better DIR [22].

The optimization time is already accelerated by cropping images around the bladder, requiring smaller meshes and reducing the granularity of the solutions. Nevertheless, this comes at the cost of a restricted field of view, which significantly affects DIR in large deformations. Hence, it is a challenge to deform an organ with a small volume in the source image that expands into a more extensive volume in the target image, where regions in the target image do not match with regions in the field of view of the source image. Moreover, the resolution of the images is reduced by half compared to the original version of MOREA, increasing the image spacing from 1.5mm to 3mm. This can have a significant impact in clinical practice as RT needs precision not to affect healthy organs, and downsampling reduces hardware requirements at the cost of a less precise result.

## 5.7 Future work

Although the proposed method shows some improvements with respect to DL-MODIR and MOREA, further research is needed. In this section, we describe some avenues for future work.

- **Evaluate on independent metrics.** MOREA and DL-MODIR evaluate the same objectives with different registration quality metrics. In this thesis, the Pareto front of the hybrid method is only compared with MOREA since both work with the same metrics, requiring a set of independent metrics for a comparison with DL-MODIR. This would permit to quantitatively compare DL and EA methods in DIR. Since each method aims to minimize a specific registration quality metric, using one method's metric to evaluate the other method would be unfair, as each is optimized for different criteria.

- **Evaluate on the original MOREA.** The MOREA method used in this thesis differs from the one in the original paper [4] in various aspects (Section 3.3.3) such as the biomechanical properties or the adaptive steering. Evaluating MOREA in the same conditions as the original paper permits a better study

of the potential benefits of the developed hybrid method and its application in clinical practice.

- **Further initial population analysis.** A more extensive analysis of the initial population permits a better understanding of the impact DL-MODIR solutions have in the hybrid method to better tune the initial population. The functional ANOVA [56] can be a strategy to evaluate the importance of different DVFs similar to the evaluation of the importance of hyperparameters in neural networks [54].

- **Develop alternative initialization strategies**. Due to time constraints, only two strategies were developed for the implemented hybrid method, including the application of the DL-MODIR DVFs with different intensities and the half-hybrid, where half of the solutions contain the DL-MODIR DVFs and the remaining half is initialized with the Gaussian attractors. Other initialization strategies can be evaluated, such as interpolating between different DL-MODIR DVF solutions to explore other regions of the approximated Pareto front and have more DVF diversity apart from tuning the intensity. Nevertheless, this would require a previous analysis of which solutions to interpolate and what the resulting DVFs look like. Also, applying Gaussian attractors on the DVFs can increase the diversity of the different DL-MODIR DVFs by adding random perturbation.

- **Further half-hybrid method analysis**. Apart from studying other methods, due to time constraints a very simple analysis is done, giving room for further research. Only the exploitation of the guidance objective is tested to evaluate it, providing a partial visualization of the method's results without knowing if solutions are more spread in the objective space or what the DVFs look like. Further analysis is needed to understand the potential benefits of the half-hybrid method since we expect that by including DL-MODIR DVFs and the Gaussian attractors, the resulting method can be comparable or better than the two methods independently.

- **Investigate other hybridization strategies**. This thesis opens the door to combining two DIR methods into a hybrid to improve each of them individually. We found a lack of research and solutions in the literature that use hybrid methods in DIR. This thesis highlights the ability and promise of hybrid methods to bridge existing gaps between individual methods. Further research combining other methods can improve existing DIR methods and bring them closer to their feasibility in some clinical practice contexts. For instance, Demons can be used to refine MOREA results as done by Huo et al. [30]. Another possibility is to provide MOREA with a higher amount of DVFs that the ones provided by DL-MODIR, which is restricted by the neural network size.

## Chapter 6

# Conclusion

This thesis proposes a hybrid method from an innovative synergy between EA-based and DL-based methods for multi-objective DIR that shows promising results compared to their former methods. DL-based methods have the advantage of a fast inference time, which can be leveraged by EA-based methods for a warm start. In this thesis, solutions from DL-MODIR are included into the initial population of MOREA at different intensities to approach more optimal DIR. The developed hybrid method demonstrated superior results for patients with large deformations between the source and target images, where other methods fall short. Specifically, MOREA cannot handle significant deformations effectively, while DL-MODIR yields less realistic deformations.

An alternative method, based on a half-hybrid strategy that initializes the population with solutions from DL-MODIR and Gaussian attractors, did not show better results than the hybrid method when exploiting the guidance objective. However, further research is needed to better understand how this method compares with the hybrid method, as well as to investigate various strategies for combining the two methods.

Despite the obstacles that the current method may face in live clinical practice due to computing limits, further analysis in real-world scenarios is necessary. However, this thesis marks a significant advancement in creating DIR methods for RT planning. Future research should focus on overcoming some of the limitations of the current method by optimizing the hardware constraints, evaluating the original MOREA version, and exploring additional hybridization strategies.

In summary, we are confident that a warm start using DL-based methods with fast inference times can significantly benefit faster convergence and better performance than EA-based methods alone in DIR. This thesis generates the foundations for more advanced and clinically viable DIR methods, with the aim to contribute to better patient outcomes in RT.

## Appendix A

# Transformed source in DL-MODIR

Solutions from DL-MODIR are included in this appendix by transforming the source contours with the provided DVFs. While some solutions get close contain small differences in the contours, others have larger differences, providing a good exploration of the objective space.

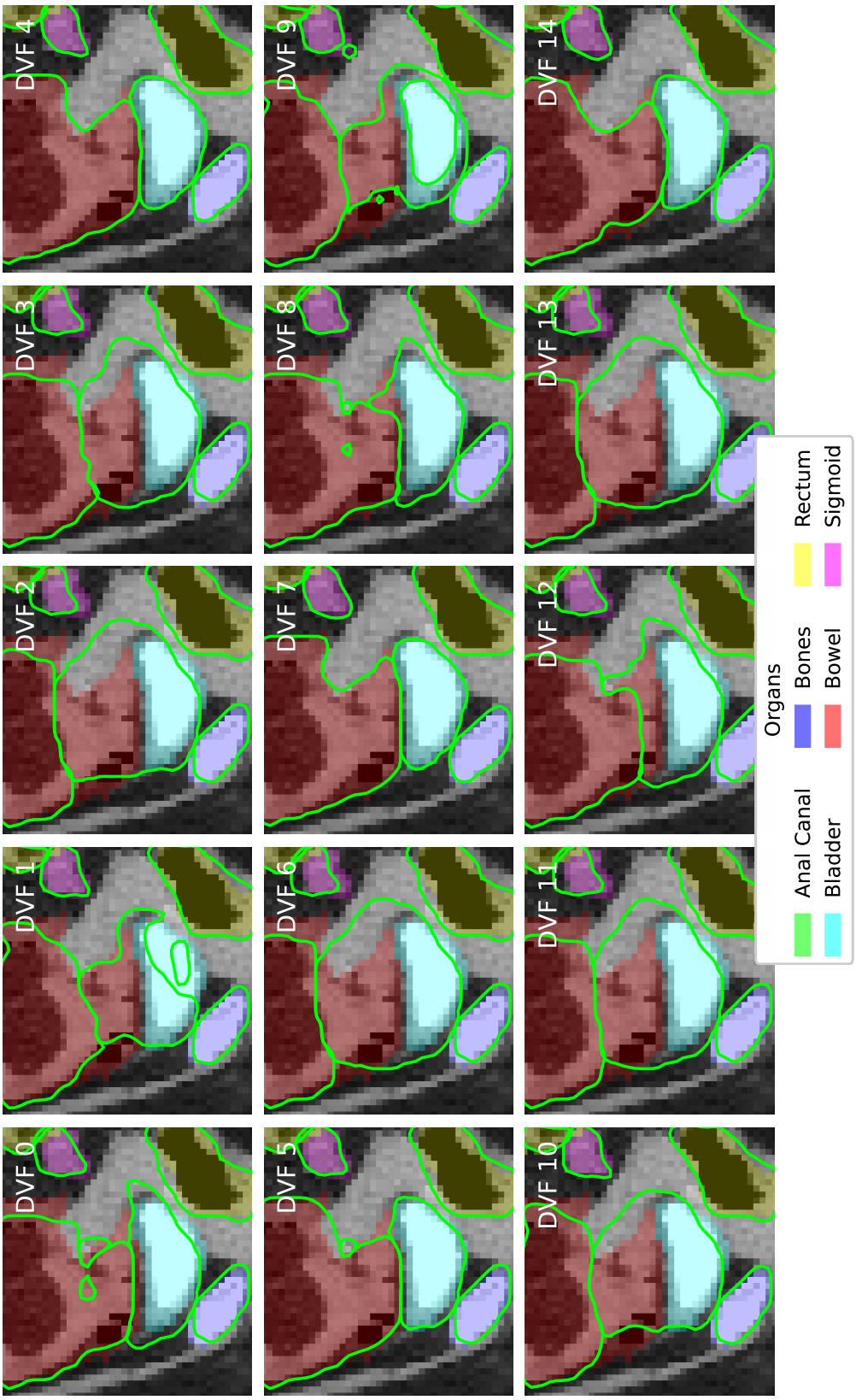



**FIGURE A.1:** Organs contours of the transformed source masks with DL-MODIR solutions in the clinical dataset.

## Appendix B

# Convergence plots

Exploitation of the guidance convergence for all patients. Patients with statistical significant difference in Table 4.2 show a bigger difference between the mean of both methods, MOREA and the hybrid method.

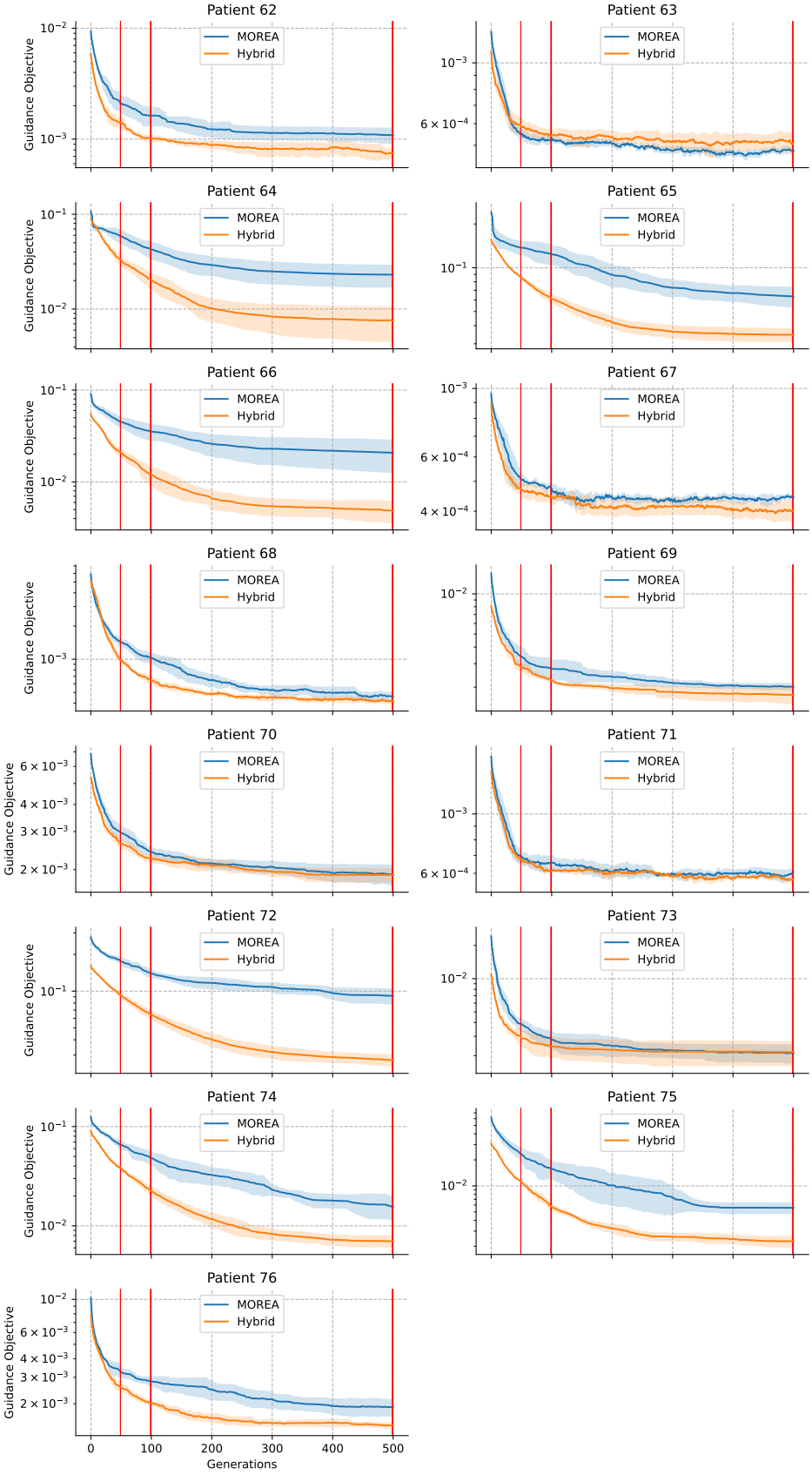



**FIGURE B.1:** Guidance convergence for the best solution along generations for all patients. Red lines represent the evaluated generations in Table 4.2 at 50, 100 and 500 generations.

## Appendix C

# Transformed source contours for all patients

This appendix includes the image of the transformed source contours of the three methods: DL-MODIR, MOREA, and the hybrid method, compared to the target image for all the patients in the sagittal axis.

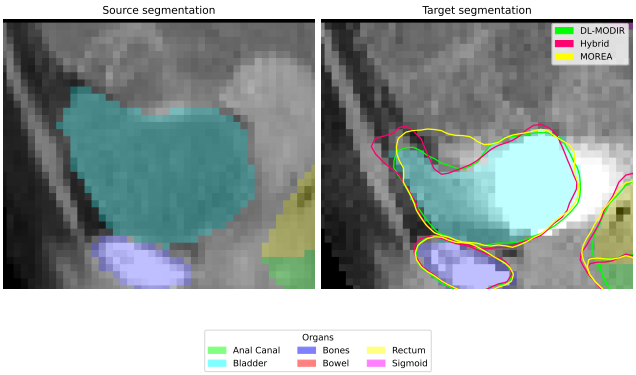



**FIGURE C.1:** Guidance contour for the three methods DL-MODIR, MOREA, and the hybrid method in patient 62.

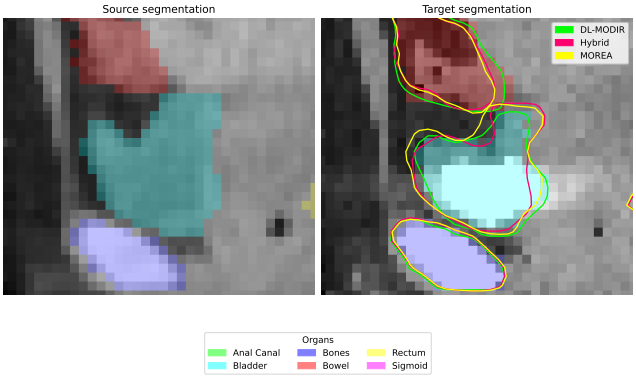



**FIGURE C.2:** Guidance contour for the three methods DL-MODIR, MOREA, and the hybrid method in patient 63.

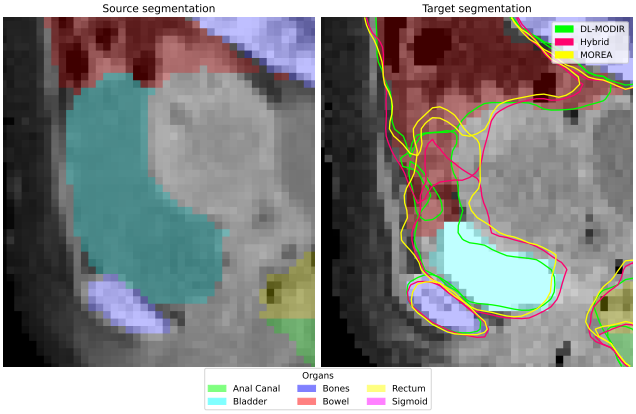



**FIGURE C.3:** Guidance contour for the three methods DL-MODIR, MOREA, and the hybrid method in patient 64.

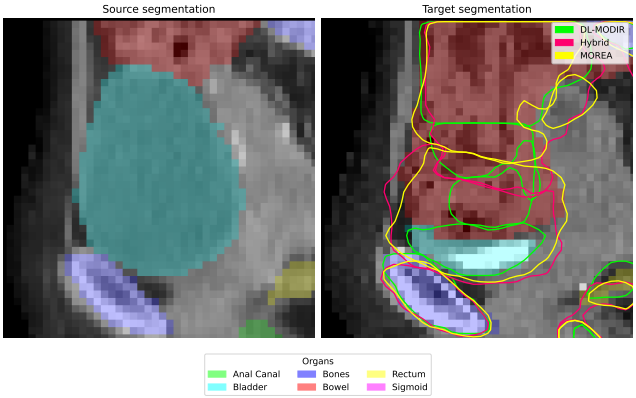



**FIGURE C.4:** Guidance contour for the three methods DL-MODIR, MOREA, and the hybrid method in patient 65.

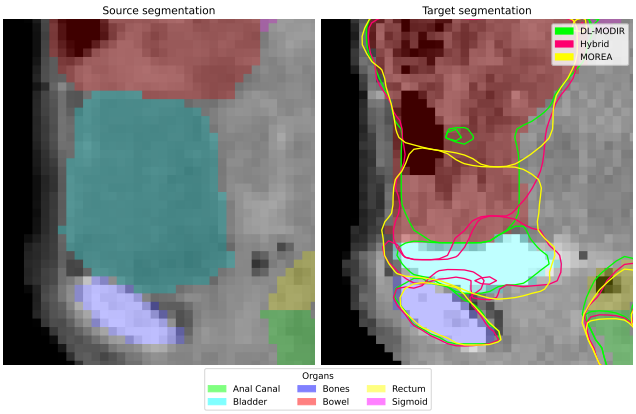



**FIGURE C.5:** Guidance contour for the three methods DL-MODIR, MOREA, and the hybrid method in patient 66.

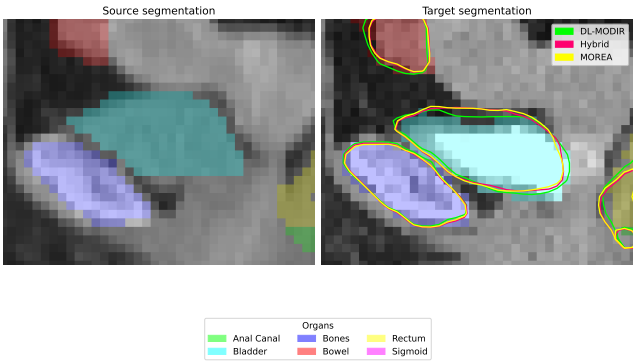



**FIGURE C.6:** Guidance contour for the three methods DL-MODIR, MOREA, and the hybrid method in patient 67.

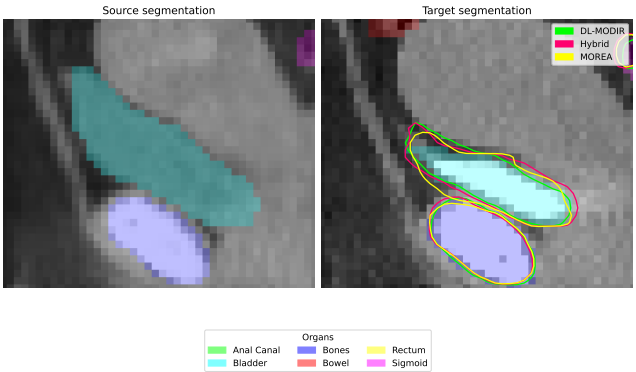



**FIGURE C.7:** Guidance contour for the three methods DL-MODIR, MOREA, and the hybrid method in patient 68.

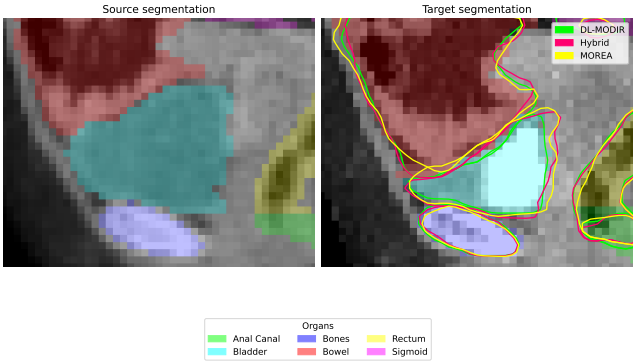



**FIGURE C.8:** Guidance contour for the three methods DL-MODIR, MOREA, and the hybrid method in patient 69.

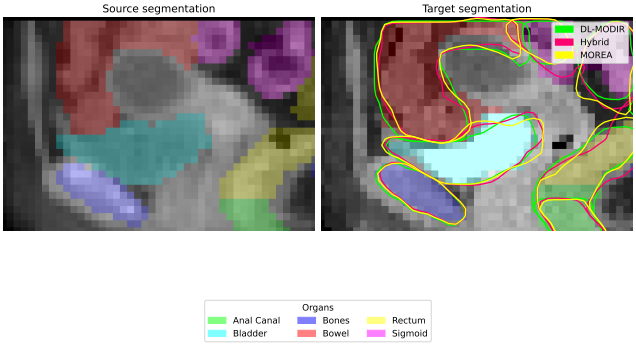



**FIGURE C.9:** Guidance contour for the three methods DL-MODIR, MOREA, and the hybrid method in patient 70.

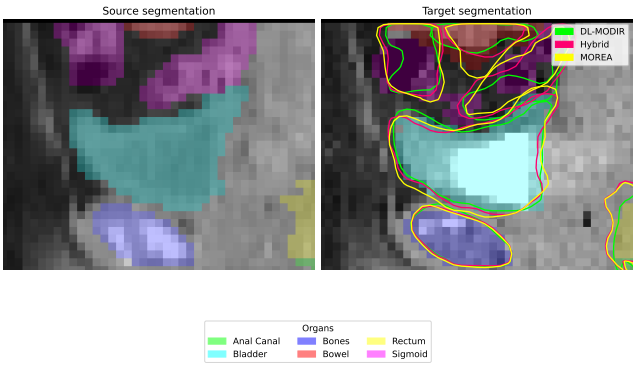



**FIGURE C.10:** Guidance contour for the three methods DL-MODIR, MOREA, and the hybrid method in patient 71.

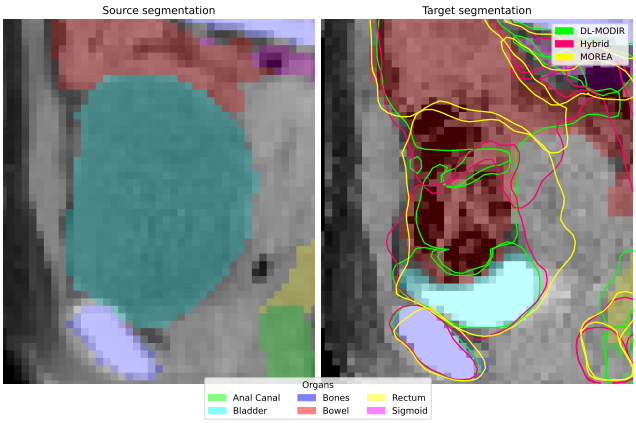



**FIGURE C.11:** Guidance contour for the three methods DL-MODIR, MOREA, and the hybrid method in patient 72.

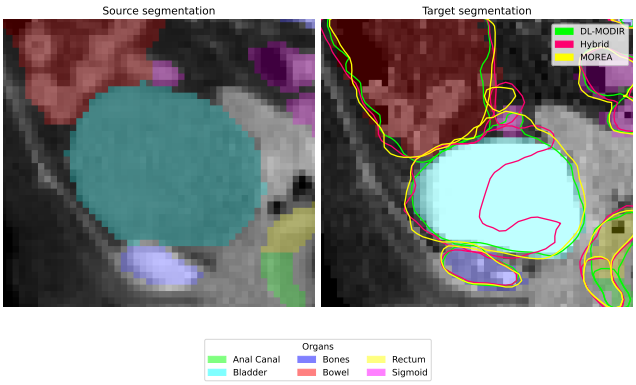



**FIGURE C.12:** Guidance contour for the three methods DL-MODIR, MOREA, and the hybrid method in patient 73.

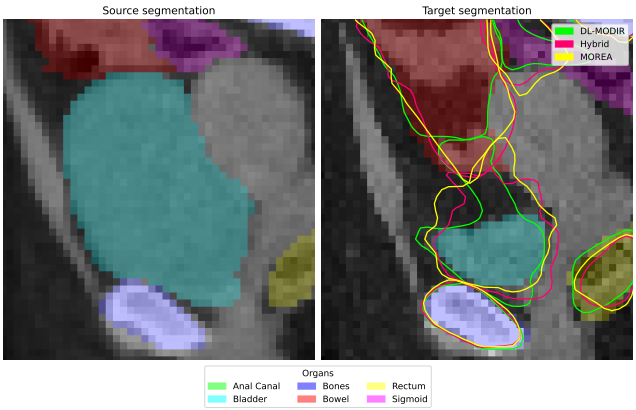



**FIGURE C.13:** Guidance contour for the three methods DL-MODIR, MOREA, and the hybrid method in patient 74.

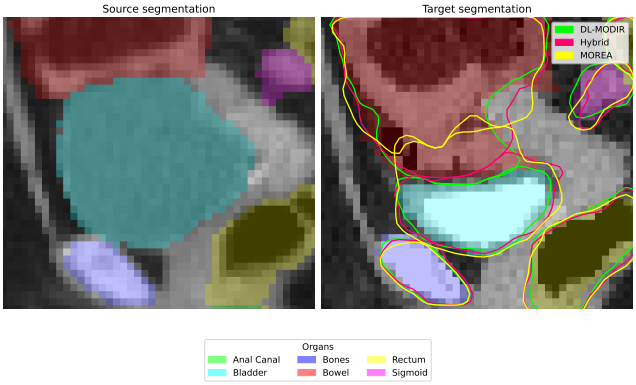



**FIGURE C.14:** Guidance contour for the three methods DL-MODIR, MOREA, and the hybrid method in patient 75.

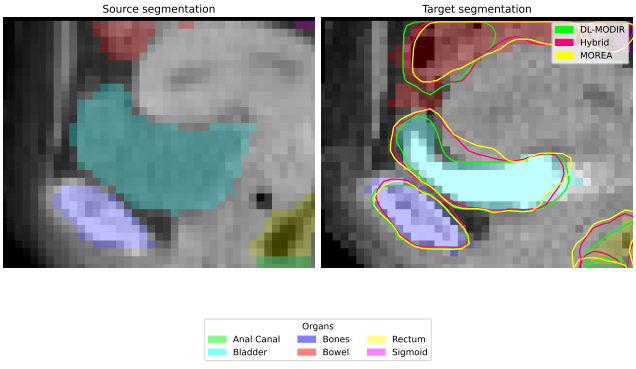



**FIGURE C.15:** Guidance contour for the three methods DL-MODIR, MOREA, and the hybrid method in patient 76.

# Bibliography

[1] Tanja Alderliesten, Jan-Jakob Sonke, and Peter A. N. Bosman. Deformable image registration by multi-objective optimization using a dual-dynamic transformation model to account for large anatomical differences. In Sebastien Ourselin and David R. Haynor, editors, *Medical Imaging 2013: Image Processing*, volume 8669, page 866910. International Society for Optics and Photonics, SPIE, 2013.

[2] N. M. Alpert, J. F. Bradshaw, D. Kennedy, and J. A. Correia. The principal axes transformation—a method for image registration. *Journal of Nuclear Medicine*, 31(10):1717–1722, 1990.

[3] Georgios Andreadis, Peter A. N. Bosman, and Tanja Alderliesten. Multi-objective dual simplex-mesh based deformable image registration for 3d medical images - proof of concept. In Ivana Išgum and Olivier Colliot, editors, *Medical Imaging 2022: Image Processing*. SPIE, April 2022.

[4] Georgios Andreadis, Peter A. N. Bosman, and Tanja Alderliesten. Morea: a gpu-accelerated evolutionary algorithm for multi-objective deformable registration of 3d medical images, 2023.

[5] Georgios Andreadis, Joas I. Mulder, Anton Bouter, Peter A. N. Bosman, and Tanja Alderliesten. A tournament of transformation models: B-spline-based vs. mesh-based multi-objective deformable image registration, 2024.

[6] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med Image Anal*, 12(1):26–41, Feb 2008.

[7] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Voxelmorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, August 2019.

[8] William A. Barrett and Eric N. Mortensen. Fast, accurate, and reproducible live-wire boundary extraction. In Karl Heinz Höhne and Ron Kikinis, editors, *Visualization in Biomedical Computing*, pages 183–192, Berlin, Heidelberg, 1996. Springer Berlin Heidelberg.

[9] Peter A.N. Bosman. The anticipated mean shift and cluster registration in mixture-based edas for multi-objective optimization. In *Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation*, GECCO '10, page 351–358, New York, NY, USA, 2010. Association for Computing Machinery.

[10] Anton Bouter, Ngoc Hoang Luong, Cees Witteveen, Tanja Alderliesten, and Peter A. N. Bosman. The multi-objective real-valued gene-pool optimal mixing evolutionary algorithm. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '17, page 537–544, New York, NY, USA, 2017. Association for Computing Machinery.

[11] K. K. Brock, M. B. Sharpe, L. A. Dawson, S. M. Kim, and D. A. Jaffray. Accuracy of finite element model-based multi-organ deformable image registration. *Med Phys*, 32(6):1647–1659, Jun 2005.

[12] Yongtao Cao, Byran J Smucker, and Timothy J Robinson. On using the hypervolume indicator to compare pareto fronts: Applications to multi-criteria optimal experimental design. *Journal of Statistical Planning and Inference*, 160:60–74, 2015.

[13] B.M. Chaparro, S. Thuillier, L.F. Menezes, P.Y. Manach, and J.V. Fernandes. Material parameters identification: Gradient-based, genetic and hybrid optimization algorithms. *Computational Materials Science*, 44(2):339–346, 2008.

[14] Indrin J Chetty and Mihaela Rosu-Bubulac. Deformable registration for dose accumulation. In *Seminars in radiation oncology*, volume 29, pages 198–208. Elsevier, 2019.

[15] Haili Chui and A. Rangarajan. A new algorithm for non-rigid point matching. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, volume 2, pages 44–51 vol.2, 2000.

[16] David Cibula, Richard Pötter, François Planchamp, Elisabeth Avall-Lundqvist, Daniela Fischerova, Christine Haie-Meder, Christhardt Köhler, Fabio Landoni, Sigurd Lax, Jacob Christian Lindegaard, et al. The european society of gynaecological oncology/european society for radiotherapy and oncology/european society of pathology guidelines for the management of patients with cervical cancer. *Virchows Archiv*, 472:919–936, 2018.

[17] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.

[18] Timo M. Deist, Stefanus C. Maree, Tanja Alderliesten, and Peter A. N. Bosman. *Multi-objective Optimization by Uncrowded Hypervolume Gradient Ascent*, page 186–200. Springer International Publishing, 2020.

[19] Institute for Health Metrics and Evaluation (IHME). Gbd 2019 cause and risk summary, 2020.

[20] Junfeng Gao, Yong Yang, Pan Lin, and Dong Sun Park. Computer vision in healthcare applications. *Journal of healthcare engineering*, 2018, 2018.

[21] J Goerres, A Uneri, M Jacobson, B Ramsay, T De Silva, M Ketcha, R Han, A Manbachi, S Vogt, G Kleinszig, et al. Planning, guidance, and quality assurance of pelvic screw placement using deformable image registration. *Physics in Medicine & Biology*, 62(23):9018, 2017.

[22] Olga L Green, Lauren E Henke, and Geoffrey D Hugo. Practical clinical workflows for online and offline adaptive radiation therapy. In *Seminars in radiation oncology*, volume 29, pages 219–227. Elsevier, 2019.

[23] Monika Grewal, Timo M. Deist, Jan Wiersma, Peter A. N. Bosman, and Tanja Alderliesten. An end-to-end deep learning approach for landmark detection and matching in medical images. In Bennett A. Landman and Ivana Išgum, editors, *Medical Imaging 2020: Image Processing*. SPIE, March 2020.

[24] Monika Grewal, Dustin van Weersel, Henrike Westerveld, Peter A. N. Bosman, and Tanja Alderliesten. Clinically acceptable segmentation of organs at risk in cervical cancer radiation treatment from clinically available annotations, 2023.

[25] Monika Grewal, Henrike Westerveld, Peter Bosman, and Tanja Alderliesten. Multi-objective learning for deformable image registration. In *Submitted to Medical Imaging with Deep Learning*, 2024. under review.

[26] Monika Grewal, Jan Wiersma, Henrike Westerveld, Peter A. N. Bosman, and Tanja Alderliesten. Automatic landmark correspondence detection in medical images with an application to deformable image registration. *Journal of Medical Imaging*, 10(01), February 2023.

[27] Peng Guan, Rebecca Howell-Jones, Ni Li, Laia Bruni, Silvia De Sanjosé, Silvia Franceschi, and Gary M Clifford. Human papillomavirus types in 115,789 hpv-positive women: a meta-analysis from cervical infection to cancer. *International journal of cancer*, 131(10):2349–2359, 2012.

[28] Kun Han, Shanlin sun, Xiangyi Yan, Chenyu You, Hao Tang, Junayed Naushad, Haoyu Ma, Deying Kong, and Xiaohui Xie. Diffeomorphic image registration with neural velocity field, 2022.

[29] Xinzi He, Jia Guo, Xuzhe Zhang, Hanwen Bi, Sarah Gerard, David Kaczka, Amin Motahari, Eric Hoffman, Joseph Reinhardt, R Graham Barr, et al. Recursive refinement network for deformable lung registration between exhale and inhale ct scans. *arXiv preprint arXiv:2106.07608*, 2021.

[30] Jinze Huo, Bangli Liu, and Qinggang Meng. A hybrid scheme for efficient deformable image registration. In Huayong Yang, Honghai Liu, Jun Zou, Zhouping Yin, Lianqing Liu, Geng Yang, Xiaoping Ouyang, and Zhiyong Wang, editors, *Intelligent Robotics and Applications*, pages 433–441, Singapore, 2023. Springer Nature Singapore.

[31] R Jadon, CA Pembroke, CL Hanna, N Palaniappan, M Evans, AE Cleves, and John Staffurth. A systematic review of organ motion and image-guided strategies in external beam radiotherapy for cervical cancer. *Clinical oncology*, 26(4):185–196, 2014.

[32] David A Jaffray and Mary K Gospodarowicz. Radiation therapy for cancer. *Cancer: disease control priorities*, 3:239–248, 2015.

[33] Mark Jenkinson and Stephen Smith. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143–156, 2001.

[34] R. Kashani, M. Hub, J. M. Balter, M. L. Kessler, L. Dong, L. Zhang, L. Xing, Y. Xie, D. Hawkes, J. A. Schnabel, J. McClelland, S. Joshi, Q. Chen, and W. Lu. Objective assessment of deformable image registration in radiotherapy: a multi-institution study. *Med Phys*, 35(12):5944–5953, Dec 2008.

[35] Theresa A Kessler. Cervical cancer: prevention and early detection. In *Seminars in oncology nursing*, volume 33, pages 172–183. Elsevier, 2017.

[36] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim. elastix: a toolbox for intensity-based medical image registration. *IEEE Trans Med Imaging*, 29(1):196–205, Jan 2010.

[37] J. Knowles and D. Corne. Properties of an adaptive archiving algorithm for storing nondominated vectors. *IEEE Transactions on Evolutionary Computation*, 7(2):100–116, 2003.

[38] Yong-xin Li, Hui Tang, Wei Wang, Xiu-feng Zhang, and Hang Qu. Dual attention network for unsupervised medical image registration based on voxelmorph. *Scientific Reports*, 12(1):16250, 2022.

[39] Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. Pareto multi-task learning. *Advances in neural information processing systems*, 32, 2019.

[40] Michalis Liontos, Anastasios Kyriazoglou, Ioannis Dimitriadis, Meletios-Athanasios Dimopoulos, and Aristotelis Bamias. Systemic therapy in cervical cancer: 30 years in review. *Critical reviews in oncology/hematology*, 137:9–17, 2019.

[41] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '87, page 163–169, New York, NY, USA, 1987. Association for Computing Machinery.

[42] Debabrata Mahapatra and Vaibhav Rajan. Multi-task learning with user preferences: Gradient descent with controlled ascent in pareto optimization. In *International Conference on Machine Learning*, pages 6597–6607. PMLR, 2020.

[43] Kaisa Miettinen. *Introduction to Multiobjective Optimization: Noninteractive Approaches*, pages 1–26. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[44] Conrado Silva Miranda and Fernando José Von Zuben. Single-solution hypervolume maximization and its use for improving generalization of neural networks. *CoRR*, abs/1602.01164, 2016.

[45] Charles A. Pelizzari, George T.Y. Chen, Danny R. Spelbring, Ralph R. Weichselbaum, and Chen Chin-Tu. Accurate three-dimensional registration of ct, pet, and/or mr images of the brain. *Journal of computer assisted tomography*, 13:20–26, 1989.

[46] K. Pirpinia, P. A. N. Bosman, C. E. Loo, N. S. Russell, M. B. van Herk, and T. Alderliesten. Simplex-based navigation tool for a posteriori selection of the preferred deformable image registration outcome from a set of trade-off

solutions obtained with multiobjective optimization for the case of breast MRI. *J Med Imaging (Bellingham)*, 5(4):045501, Oct 2018.

[47] K. Pirpinia, P. A. N. Bosman, C. E. Loo, G. Winter-Warnars, N. N. Y. Janssen, A. N. Scholten, J. J. Sonke, M. van Herk, and T. Alderliesten. The feasibility of manual parameter tuning for deformable breast MR image registration from a multi-objective optimization perspective. *Phys Med Biol*, 62(14):5723–5743, Jun 2017.

[48] A. Ramkumar, J. Dolz, H. A. Kirisli, S. Adebahr, T. Schimek-Jasch, U. Nestle, L. Massoptier, E. Varga, P. J. Stappers, W. J. Niessen, and Y. Song. User Interaction in Semi-Automatic Segmentation of Organs at Risk: a Case Study in Radiotherapy. *J Digit Imaging*, 29(2):264–277, Apr 2016.

[49] Shayna E. Rich. 7 - selecting modalities. In Neha Vapiwala, Joshua A. Jones, and Kavita V. Dharmarajan, editors, *Palliative Radiation Oncology*, pages 55–64. Elsevier, Philadelphia, 2024.

[50] B Rigaud, A Klopp, S Vedam, A Venkatesan, N Taku, A Simon, P Haigron, R de Crevoisier, K K Brock, and G Cazoulat. Deformable image registration for dose mapping between external beam radiotherapy and brachytherapy images of cervical cancer. *Physics in Medicine  Biology*, 64(11):115023, may 2019.

[51] Petter Risholm, Alexandra J Golby, and William Wells. Multimodal image registration for preoperative planning and image-guided neurosurgical procedures. *Neurosurgery Clinics*, 22(2):197–206, 2011.

[52] Cedric Rodriguez, Stephanie de Boer, Peter Bosman, and Tanja Alderliesten. Bi-objective optimization of organ properties for the simulation of intracavitary brachytherapy applicator placement in cervical cancer. Number 124660H, April 2023.

[53] D. Rueckert, L.I. Sonoda, C. Hayes, D.L.G. Hill, M.O. Leach, and D.J. Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, 1999.

[54] Abhinav Sharma, Jan N van Rijn, Frank Hutter, and Andreas Müller. Hyperparameter importance for image classification by residual neural networks. In *Discovery Science: 22nd International Conference, DS 2019, Split, Croatia, October 28–30, 2019, Proceedings 22*, pages 112–126. Springer, 2019.

[55] Hang Si. Tetgen, a delaunay-based quality tetrahedral mesh generator. *ACM Transactions on Mathematical Software*, 41(2):Article 11, February 2015.

[56] I.M. Sobol. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 4:407–414, 1993.

[57] Hessam Sokooti, Bob de Vos, Floris Berendsen, Mohsen Ghafoorian, Sahar Yousefi, Boudewijn P. F. Lelieveldt, Ivana Isgum, and Marius Staring. 3d convolutional neural networks image registration based on efficient supervised learning from artificial deformations, 2019.

[58] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. Deformable medical image registration: A survey. *IEEE Transactions on Medical Imaging*, 32(7):1153–1190, 2013.

[59] Marius Staring, Stefan Klein, and Josien P. W. Pluim. A rigidity penalty term for nonrigid registration. *Medical Physics*, 34(11):4098–4108, 2007.

[60] Kenji Suzuki. Overview of deep learning in medical imaging. *Radiological physics and technology*, 10(3):257–273, 2017.

[61] Jan-Jakob Sonke Tanja Alderliesten and Peter A. N. Bosman. Deformable image registration by multi-objective optimization using a dual-dynamic transformation model to account for large anatomical differences. In Sebastien Ourselin and David R. Haynor, editors, *Medical Imaging 2013: Image Processing*, volume 8669, page 866910. International Society for Optics and Photonics, SPIE, 2013.

[62] J.P. Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical Image Analysis*, 2(3):243–260, 1998.

[63] Jan Unkelbach, Markus Alber, Mark Bangert, Rasmus Bokrantz, Timothy CY Chan, Joseph O Deasy, Albin Fredriksson, Bram L Gorissen, Marcel Van Herk, Wei Liu, et al. Robust radiotherapy planning. *Physics in Medicine & Biology*, 63(22):22TR02, 2018.

[64] Maureen van Eijnatten, Leonardo Rundo, K Joost Batenburg, Felix Lucka, Emma Beddowes, Carlos Caldas, Ferdia A Gallagher, Evis Sala, Carola-Bibiane Schönlieb, and Ramona Woitek. 3d deformable registration of longitudinal abdominopelvic ct images using unsupervised deep learning. *Computer Methods and Programs in Biomedicine*, 208:106261, 2021.

[65] Diana Veiga-Canuto, Leonor Cerdà-Alberich, Cinta Sangüesa Nebot, Blanca Martínez de las Heras, Ulrike Pötschger, Michela Gabelloni, José Miguel Carot Sierra, Sabine Taschner-Mandl, Vanessa Düster, Adela Cañete, et al. Comparative multicentric evaluation of inter-observer variability in manual and automatic segmentation of neuroblastic tumors in magnetic resonance images. *Cancers*, 14(15):3648, 2022.

[66] Grace Wahba. *Spline Models for Observational Data*. SIAM, 1990.

[67] G. A. Whitfield, P. Price, G. J. Price, and C. J. Moore. Automated delineation of radiotherapy volumes: are we going in the right direction? *Br J Radiol*, 86(1021):20110718, Jan 2013.

[68] S. Wognum, L. Bondar, A. G. Zolnay, X. Chai, M. C. Hulshof, M. S. Hoogeman, and A. Bel. Control over structure-specific flexibility improves anatomical accuracy for point-based deformable registration in bladder cancer radiotherapy. *Med Phys*, 40(2):021702, Feb 2013.

[69] World Health Organization. Cervical cancer, 2024.

[70] Feng Zhao, Qingming Huang, and Wen Gao. Image matching by normalized cross-correlation. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 2, pages II–II, 2006.

[71] H. Zhong, J. Kim, and I. J. Chetty. Analysis of deformable image registration accuracy using computational modeling. *Med Phys*, 37(3):970–979, Mar 2010.

[72] Hualiang Zhong, Jinkoo Kim, Haisen Li, Teamour Nurushev, Benjamin Movsas, and Indrin J. Chetty. A finite element method to correct deformable image registration errors in low-contrast regions. *Physics in Medicine and Biology*, 57:3499, 6 2012.