UCLouvain

epl

École polytechnique de Louvain

Trustworthy AI : Second opinion needed for faster and safer clinical deployment of AI

Author: **Arnaud Bacq**
Supervisor: **Benoit Macq**
Readers: **John A. Lee, Dani Manjah, Jean Léger**
Academic year 2021–2022
Master [120] in Computer Science and Engineering

Acknowledgements

My sincere thanks go to Professor Benoit Macq for his supervision. Our meetings were an opportunity to explore new ideas and think deeply about the objectives of this master thesis.
I would like to thank Dani Manjah for his guidance during this master thesis, his helpful advice and constructive criticism were essential to this work.
I want to thank Jean Léger for his availability during the first months of this work. He gave me the opportunity to work on real data and shared with me valuable programming suggestions. I really appreciated that he always took the time to answer my numerous emails and to guide me on the right path.
I appreciate that Maxime Zanella took the time to share his knowledge of active learning, his expertise was a great asset.
I would like to thank Brigitte Dupont and François Hubin for the opportunity to perform all the computation remotely on a server of the university.
Many thanks to my friends and family for being a constant source of support.
Finally, I want to thank Andréa, I'm truly grateful for your kindness and understanding.

Abstract

The development of fast and reliable tools for medical image segmentation is necessary to meet the ever-growing demand in radiography. The models showing the best performance are based on deep neural networks but necessitate a sufficiently large and diverse annotated dataset. This is often unavailable due to the high level of expertise required to annotate the data and privacy regulations. We propose a multi-method and active learning approach to reduce the amount of annotated data needed for training. It compares the predictions of a U-Net model and a Morphons algorithm following a query-by-committee strategy to specifically label the most informative images to add to the training set. Our approach was tested against random selection for bladder and rectum segmentation. We also propose an intelligent system relying on atlas-based segmentation when the limited size of the dataset hinders the performance of the deep learning model. Thus, we show that atlas-based segmentation is a useful second opinion for deep learning models to reduce the amount of annotated data needed and maintain good performance when working with small datasets.

Contents

List of Figures

List of Tables

Abbreviations

ABAS Atlas-based auto-segmentation

Adam Adaptive moment estimation

AI Artificial Intelligence

ANN Artificial Neural Network

CBCT Cone Beam Computed Tomography

CNN Convolution neural network

CT Computed Tomography

DL Deep Learning

DNN Deep Neural Network

DSC Dice similarity coeficient

FCN Fully convolutional network

FCNN Fully convolutional neural network

FN False Negative

FNN Feedforward Network

FP False Positive

IoU Intersection over Union

JI Jaccard Index

MAS Multi-atlas segmentation

ML Machine Learning

MLP Multilayer Perceptron

MRI Magnetic Resonance Imaging

NN Neural Network

OAR Organs at Risk

ReLU Rectified linear unit

ROI Region of interest

SOA State of the art

TN True Negative

TP True Positive

Chapter 1

Introduction

Worldwide, nearly 10 million people died from cancer in 2020, this represents nearly one in six death [21]. The same year, in 2020, the world faced about 19.3 million new cancer cases and the number of new cases is expected to reach 28.4 million in 2040 [22]. Therefore, it becomes increasingly important to develop new ways to treat patients as safely and efficiently as possible.

To solve this issue, many cancer treatments have been developed: chemotherapy, immunotherapy, surgery, radiation therapy,... Among them, radiation therapy plays a significant role, indeed external beam radiotherapy is indicated in around 52% of cases [23]. Radiation therapy involves using high doses of radiation to kill cancer cells or slow their growth by damaging their DNA. Compared to chemotherapy, this method has the benefit of not exposing the entire body to the treatment by targeting only the part of the body where the cancer is located.

The recent development in Artificial Intelligence (AI) has opened new paths for improvement in every step of the radiation therapy workflow, from the initial treatment decision to follow-up care [1].

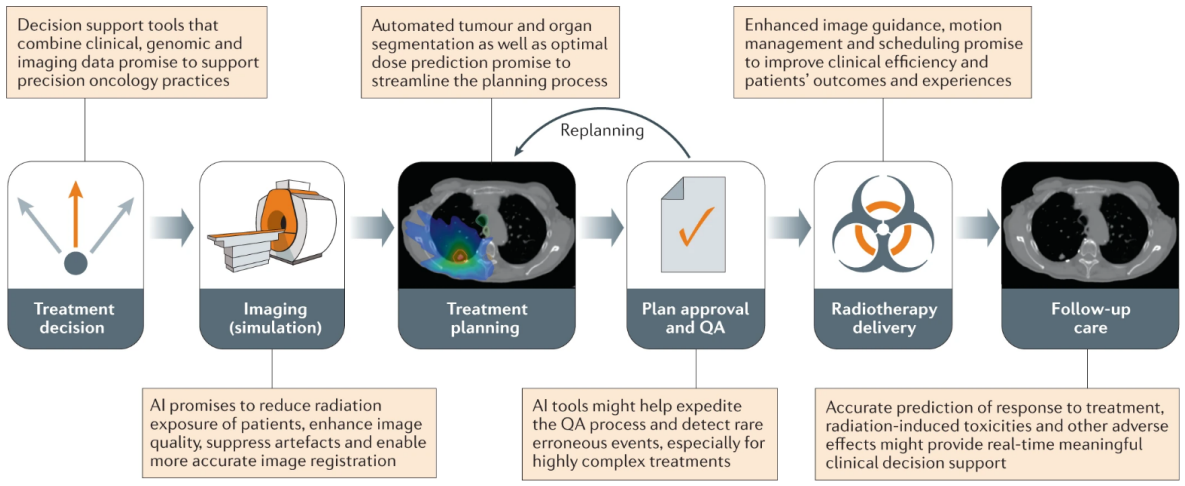


Figure 1.1: Radiation therapy workflow highlighting paths of improvement using AI. [1].

A major concern during treatment planning is to establish the precise location of the tumor and delineate its boundaries. This process, called medical image segmentation [24], is also performed on the organs adjacent to the tumor in order to calculate the radiation dose delivered to these critical organs and ensure that it remains within safe limits [1].

Manual segmentation is performed by a radiation oncologist, usually on Magnetic Resonance Imaging (MRI), Computed Tomography (CT), or Cone Beam Computed Tomography (CBCT) scans, it is a very time-consuming process that requires a high level of expertise and must be done very carefully because inaccuracies in the segmentation lead to over- or underdosing of the tumor or surrounding healthy organs [4].

Automated segmentation techniques have been developed in an attempt to solve this problem with State of the art (SOA) techniques such as atlas-based segmentation, a method based on image registration, obtaining robust results and being used for the segmentation of abdominal fat [25], the heart wall [8], lung tumor [26], the prostate in prostate cancer [27] and many others. Recently, the development of machine learning and deep learning has opened up new perspectives for automated image segmentation. For example, the use of CNN such as the U-Net architecture has already shown great results, outperforming other state-of-the art segmentation techniques on a wide variety of segmentation tasks in medical imaging [19].

Despite good performance, the deployment of AI and Deep Learning (DL) in the medical field comes with some challenges. The need for large amounts of labeled data to train deep learning models is a real problem for their clinical deployment. For example, medical image segmentation typically faces limited datasets [28]. Deep Learning models are also criticized for their lack of interpretability. Indeed, the complexity of the models and the large number of hidden layers make it difficult for a human to understand how they reach their conclusion, which leads to their use as a "blackbox" [29].

This master thesis proposes an approach in active learning, a machine learning paradigm used to reduce the amount of labeled data needed for training of the model [20]. The approach combines both deep learning and atlas-based segmentation following a "Query-By-Committee" strategy that is tested on real data for bladder and rectum segmentation. At the end of this master thesis, we propose an intelligent system that relies on atlas-based segmentation when the training set is too small and switches to deep learning when the size of the training set becomes sufficient. This system aims to achieve the best possible performance in all environments, even when the amount of available training data is too small for a deep learning model to achieve good performance.

This master thesis is divided into 4 parts. First, we discuss the state of the art of segmentation techniques in medical imaging and the state of the art of active learning with a particular focus on the query-by-committee strategy. The next chapter explains our proposed approach to reduce the amount of labeled data required. Chapter 4 presents our results, the experimental methodology we followed, and contains a discussion of our results and the limitations of our approach. Finally, in the chapter Future work, we propose an intelligent system relying on atlas-based segmentation when the limited size of the dataset hinders the performance of the deep learning model.

Chapter 2

State of the art

In this chapter, we discuss the state of the art of segmentation techniques in medical imaging. Segmentation is crucial in dose planning of radiotherapy sessions, it establishes the precise location of the tumor and Organs at Risk (OAR) and delineates their boundaries. We start with atlas-based segmentation, a segmentation technique based on image registration that extracts prior knowledge from a reference image that we call an atlas [30]. Next, we explain segmentation techniques that use active contour (also known as SNAKES), a method that formulates contouring as an optimization problem [20]. Finally, we describe the development of convolutional neural networks in medical imaging and the use of fully convolutional networks like the U-Net architecture for segmentation tasks [19].

Then, we discuss the state of the art of active learning, a machine learning paradigm that attempts to reduce the amount of labeled data needed to train a machine learning model by choosing the images that will improve the model the most [20]. Finally, we briefly discuss the performance measurements of segmentation tasks in medical imaging.

2.1 Segmentation techniques

CT segmentation

Segmentation is a crucial task in dose planning for radiotherapy sessions within the radiation therapy workflow. It has to be done carefully to minimize radiation exposure to the organs at risk and avoid any complications for the patient. Planning is often done on CT scans, but manual annotation of the data requires a high level of expertise and is very time-consuming for the radiologist. This problem resulted in the development of automated image segmentation. Figure[2.1] shows different segmentations techniques that can be used on CT scans [2].

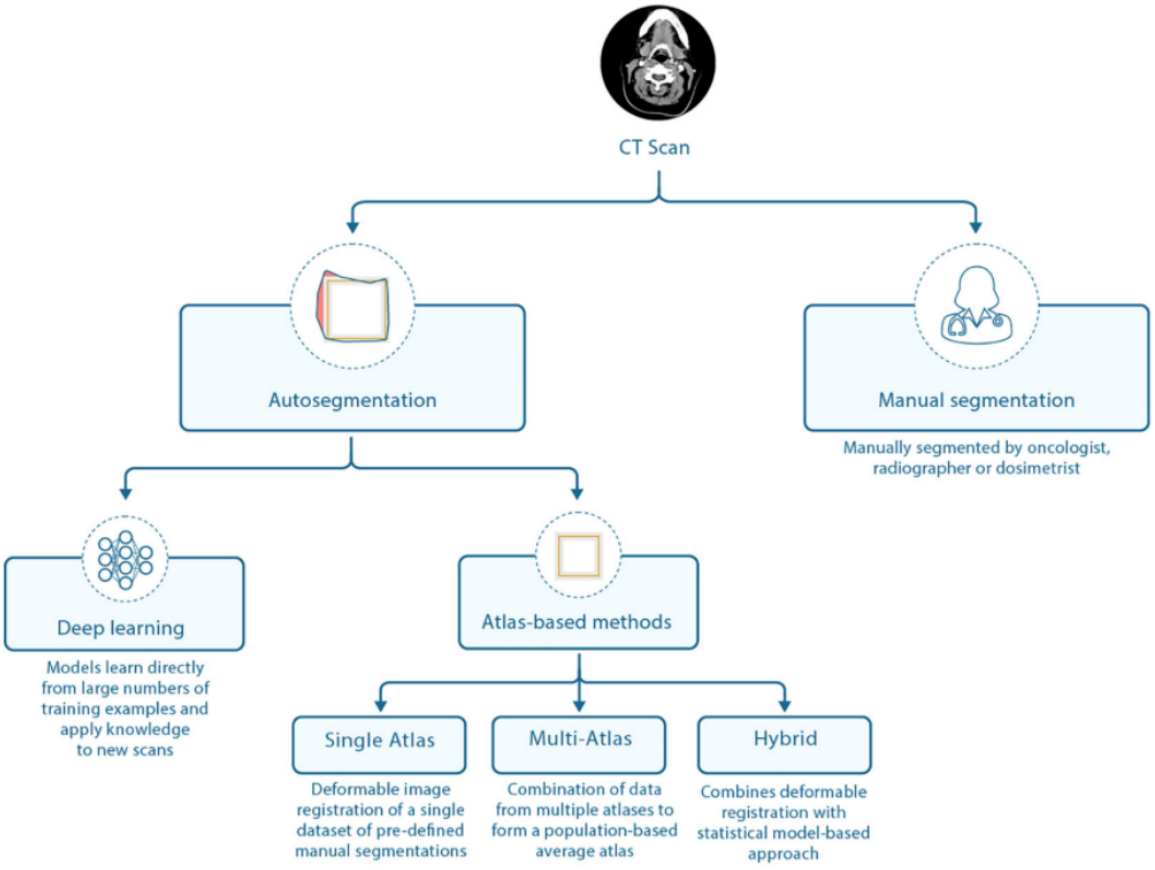


Figure 2.1: Different Segmentation techniques used on CT scans [2].

CBCT segmentation

In radiotherapy, we observe large anatomical deformations between planning and treatment sessions, caused by tumor growth/shrinkage, weight loss or internal movements [31] [32]. Anatomical changes are particularly common in the pelvic region due to bladder and rectal movements [33], leading to uncertainty in the dose delivered to the tumor and surrounding healthy organs [5]. Figure[2.2] gives an example in a lung cancer case where the tumor shrinkage between the planning CT and the treatment session (5 weeks later) induces a beam overshoot [3] [4].

(a) Planning CT (b) CT after 5 weeks

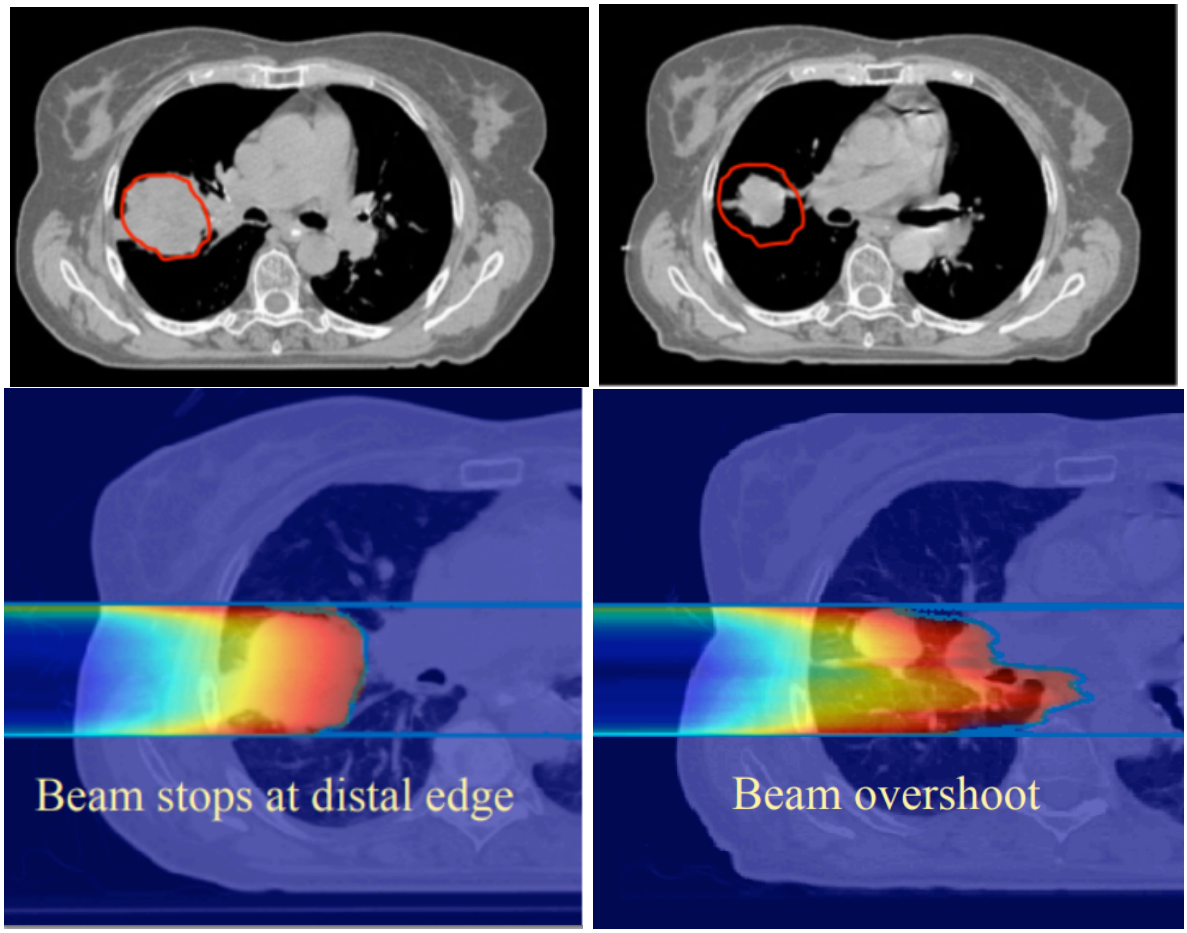
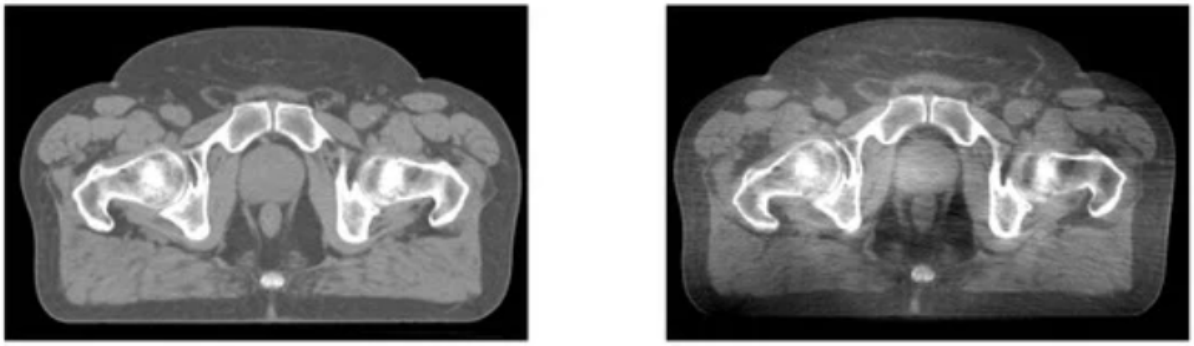


Figure 2.2: Impact of tumor shrinkage between planning and radiation exposure [3][4].

CBCT scans are performed on the day of treatment to monitor these deformations because it tends to be easier than repeating a CT scan which requires a separate appointment and results in additional burden, cost, and radiation exposure for the patient [34]. But working with CBCT scans can be challenging because they are of lower quality than CT scans due to various artifacts including noises, beam hardening, and scattering [5]. Figure[2.3] shows a comparison of a CT and a CBCT scan of the pelvis.

This section develops 3 different segmentation techniques that can be used. Atlas-based segmentation is a very popular method based on image registration that extracts prior knowledge from a reference image that we call an atlas [30], with different variants using on single or multi-atlas segmentation. Segmentation techniques using active contour are methods that formulate contouring as an optimization problem [20]. Finally, Deep Learning techniques are the most recent techniques but have already shown great results managing to outperform the other techniques on a wide variety of segmentation tasks [19].



(a) Slice of a CT scan.

(b) Slice of a CBCT scan.

Figure 2.3: Comparison of a CT and a CBCT scan of the pelvis [5].

2.1.1 Atlas-based segmentation

Most people share anatomical similarities, for example, the heart placed slightly on the left of the chest, etc. Combining many delineated images of patients sharing anatomical similarities (e.g. patients of the same sex), we can define a delineated image called an atlas representing the segmentation of an "average" patient. The assumption made, is that the similarities are sufficient to be able to deform the body of someone into the body of someone else [35]. The atlas is essentially a reference image used for the deformation, for example, a labeled CT scan can be used to label a CBCT scan.

Image registration is simply the calculation of the deformation from one image to another. If we use an atlas with image registration, we obtain a segmentation technique called atlas-based segmentation, also called Atlas-based auto-segmentation (ABAS). ABAS extracts prior knowledge from the atlas [30] by performing image registration between the image we query and the atlas to map their space coordinates through a series of transformations and deformations [36].

Image registration

Image registration can be used to label a new image with a reference image or to monitor deformations of the OARs between planning and treatment sessions. Figure[2.4] shows an illustration of image registration.

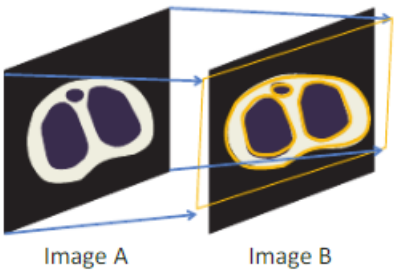


Figure 2.4: Illustration of image registration [6].

To map the different images, the image can be deformed and transformed with different transformations possible that we call geometric transformations [37]. We distinguish two categories of transformations possible linear and non-linear transformations. Linear transformations are composed of translation, rotation, scaling, shear, and affine transformations being an arbitrary combination of the prior transformations [6]. Non-linear transformations are also possible, they have the particularity to map straight lines to a curve, whereas linear transformations map any straight line to a straight line. This typically happens when the images are from different modalities (e.g. CT and MRI) [35]. Figure[2.5] shows the geometric transformation functions for image registration.

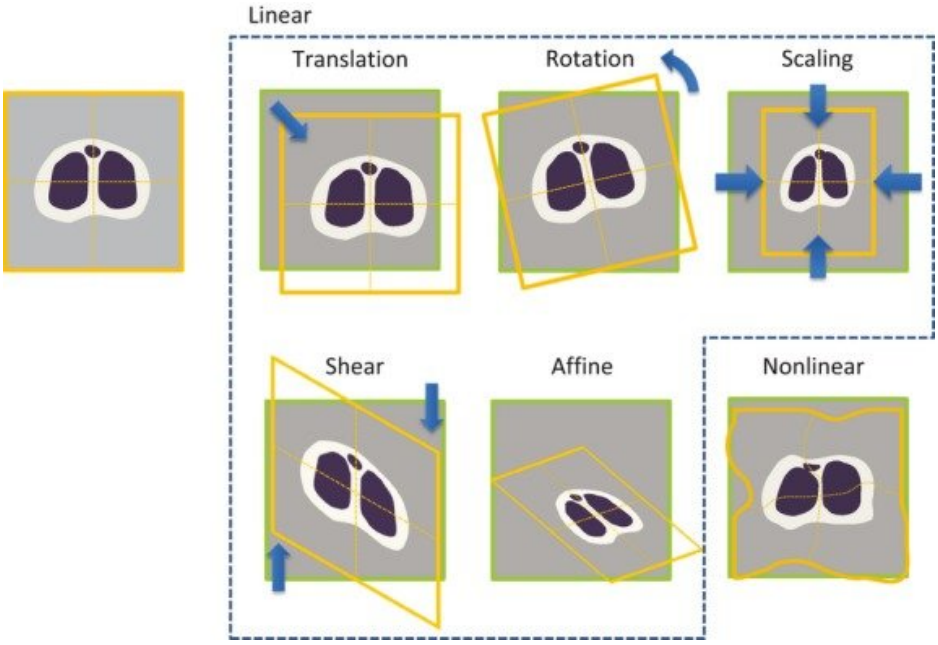


Figure 2.5: Geometric transformation functions for image registration, composed of linear and nonlinear transformations [6].

In the literature, rigid registration means that we achieve correspondence simply with translations and rotations. To monitor anatomical deformations, non-rigid registration is often required. To implement non-registration, vector fields, called deformation fields, sharing the same sampling grid between the two images are often used [35]. If the model has a small set of parameters describing the deformation, the task is called parametric registration [38]. Otherwise, when each pixel has its own unconstrained deformation vector, which can be optimized independently of its neighbors, it is called non-parametric registration [35].

The Morphons algorithm

The Morphons algorithm is a non-parametric image registration algorithm. A particularity of the Morphons algorithm is its metric: it computes the local phase for each voxel of the image to perform the displacement estimation. The local phase of each voxel is computed by convolving both images with directional quadrature filters [39]. Quadrature filters typically contain a real part to detect the lines and an imaginary part to detect the edges [7]. Figure[2.6] shows the main concept of the local phase.

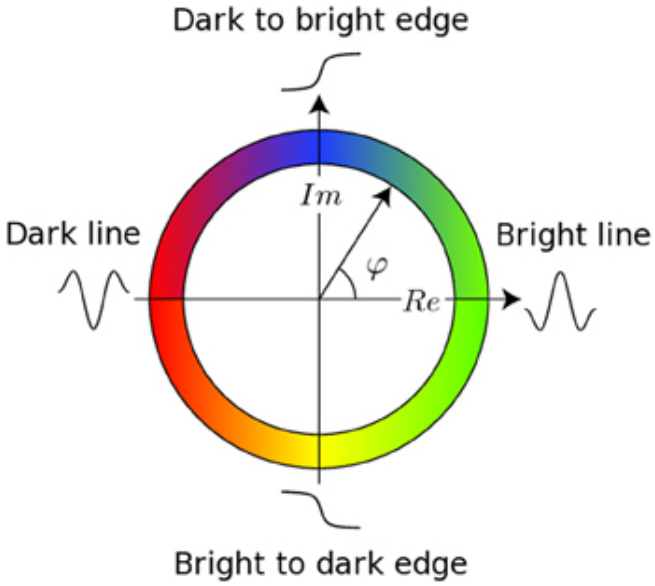


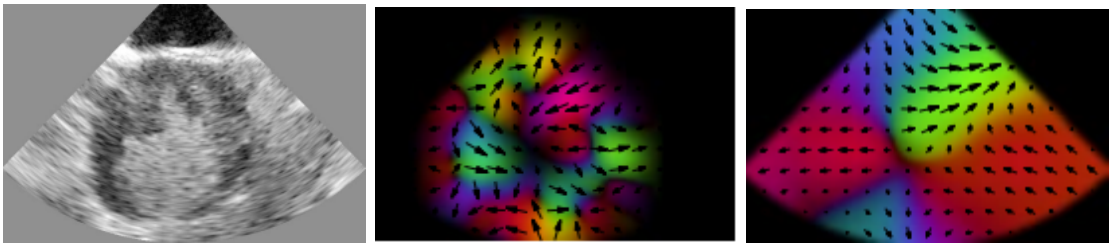
Figure 2.6: Illustration of the phase from a quadrature response [7]

To be able to find a large and accurate deformation, the Morphons algorithm follows a multi-resolution approach: at each resolution level, the images and deformation field are re-sampled to have a specific size [39]. The Morphons algorithm is, therefore, an iterative process where the deformation field is accumulated at each resolution level. The deformation process can be divided into three steps: displacement estimation, deformation field accumulation, and deformation.

- The displacement estimation is computed using the local phase difference in the two images.
- An incremental displacement field is calculated according to the new displacement estimation at this resolution level. The accumulated deformation field is updated with the incremental displacement field. Smoothing is also applied to enforce the smoothness of the deformation field.
- The moving image is deformed according to the updated accumulated deformation field.

In the first paper presenting the Morphons algorithm [8], Hans Knutsson and Mats Andersson have shown a representation of the incremental displacement field and the accumulated displacement field as they were used on 2D ultrasound images of a heart to segment the heart wall (Figure[2.7]).

After being used for the first time to segment the heart wall [8], the Morphons algorithm has shown good results in many other applications. For the segmentation of abdominal fat on MRI [25], 2D photographs of hands and faces, 3D CT data of the hip region, and 3D MR brain images [40], or on CT scan for lung cancer [26].



(a) Ultrasound image of a heart (b) Incremental displacement field (c) Accumulated displacement field

Figure 2.7: Representation of the incremental displacement field and the accumulated displacement field used on a 2D ultrasound image of a heart [8].

The Morphons registration is still evolving, a diffeomorphic version of the Morphons algorithm has been proposed in [41]. Diffeomorphism ensures that the displacement field is physical [42], this is a desirable property because organs can be compressed and deformed but cannot undergo noninvertible spatial transformations [41].

The Demons algorithm is another non-parametric image registration algorithm worth mentioning. It uses a method based on the calculation of a displacement field to match the intensities in both images [43].

Multi-atlas segmentation

Multi-atlas segmentation (MAS) was introduced to improve single atlas-based segmentation, the idea is to make more informed decisions for the label of the input image by taking advantage of many atlases. By using several atlases, MAS wants to offer a better segmentation accuracy by being more informed on the possible anatomical variations. However, manipulating more atlases also comes at the price of a higher computational cost [44].

The core component of MAS methods is the combination of the different atlas labels called label fusion. The most simple approach for label fusion is to select the best atlas by comparing image intensities between the atlases and the input image [45]. However, relying on a single atlas is not ideal and pretty much comes back to single atlas segmentation. Therefore, methods based on majority voting have been developped to keep the useful information in all atlases, for example choosing the most frequent label at each location [46].

An interesting approach for label fusion is the "Simultaneous truth and performance level estimation" (STAPLE) algorithm which was first introduced to compare different manual segmentations by experts, using them to get closer to the hidden, ground-truth segmentation [47]. Compared to majority voting, STAPLE does not consider that all atlases perform equally on the target image. Instead, the algorithm evaluates the performance of all atlases, incorporate them inside a probabilistic framework and finds an optimal combination, weighting each segmentation based on their estimated performance level. This approach to label fusion has proven more robust to anatomical variation than majority voting [47].

2.1.2 Active contour

The active contour model is a method used to find a closed contour of a target object and can therefore be used for segmentation. The key idea of the model was to formulate the contour detection problem as an optimization method, finding the best contour by optimizing a certain criterion, for example, the smoothness of the contour [6]. The method is also called SNAKES in reference to the use of energy-minimizing splines inside the original model proposed by Kass et al. Figure[2.8] shows how the snake is guided by constraints inside the model to identify nearby edges, lines and curves [9].

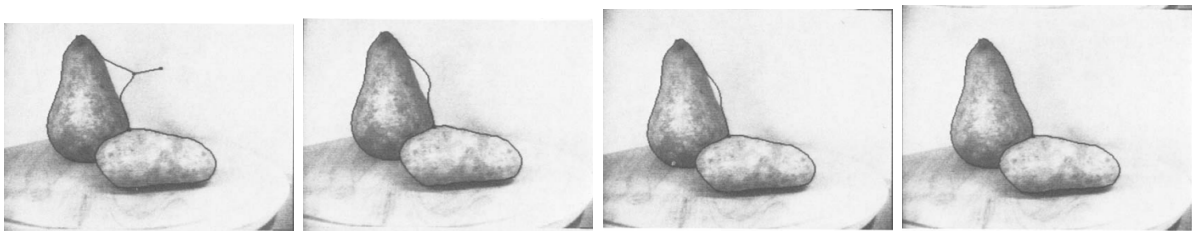


Figure 2.8: Representation of a snake converging from its initial position away of the pear (left) to the contour of the pear (right) [9].

This method has shown good results for tumor segmentation on ultrasound images [10]. Figure[2.9] shows the initial contour chosen and the boundary derived by the method. A recurrent problem was that the initial contour had to be close enough to the solution for the algorithm to converge. To solve this issue, a modified architecture has been proposed in [48] that changes the forces pushing the curve to the edge to obtain more stable results. This architecture allows the curve to pass over weak edges and stop only if the edge is strong, reducing the need for the initial curve to be close to the solution.

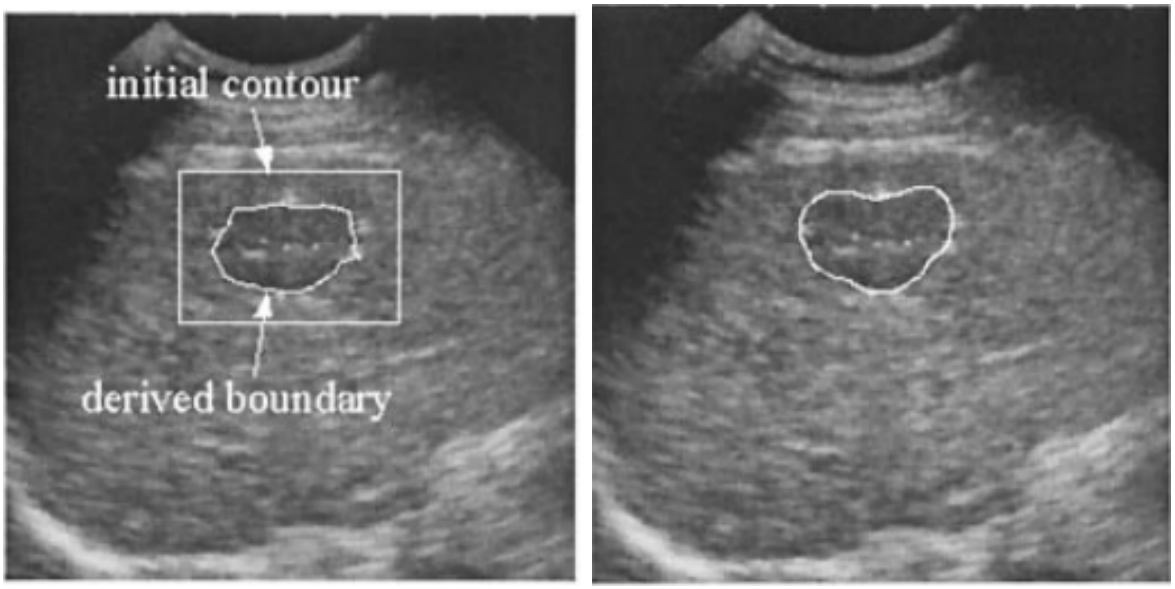


Figure 2.9: Snakes model used for tumor segmentation on ultrasound images on the left and the manual segmentation on the right [10].

2.1.3 Convolutional neural networks

In 1988, the first Deep Neural Network (DNN) using convolutions named the neocognitron was proposed [49] but at the time these networks didn't get much attention, among other things, due to the lack of computational power. In 1998 Lecun et al. proposed to apply a gradient-based learning algorithm to CNN, creating the LeNet architecture and opening the path for the CNN we know today [50].

In 2012 a CNN named AlexNet won the most difficult ImageNet challenge for visual object recognition [51]. The large ImageNet dataset with millions of annotated data was perfect for the deployment of deep learning and the network achieved significant results against all traditional machine learning and computer vision approaches. This was a major turning point for Deep Learning and helped to raise a lot of interest for these networks. CNNs continue to gather a lot of attention in medical image segmentation because they outperform other state-of-the-art on a wide variety of segmentation tasks [52].

CNN Architecture

CNNs can be classified as a subpart of DNNs, their name comes from the convolutional layers that constitute their architectures. Figure[2.10] shows the overall architecture of a CNN. It can be divided into two parts: feature extraction and classification. Feature extraction is performed by multiple alternating convolution and max-pooling layers. After feature extraction, the data is classified by fully connected layers [11].

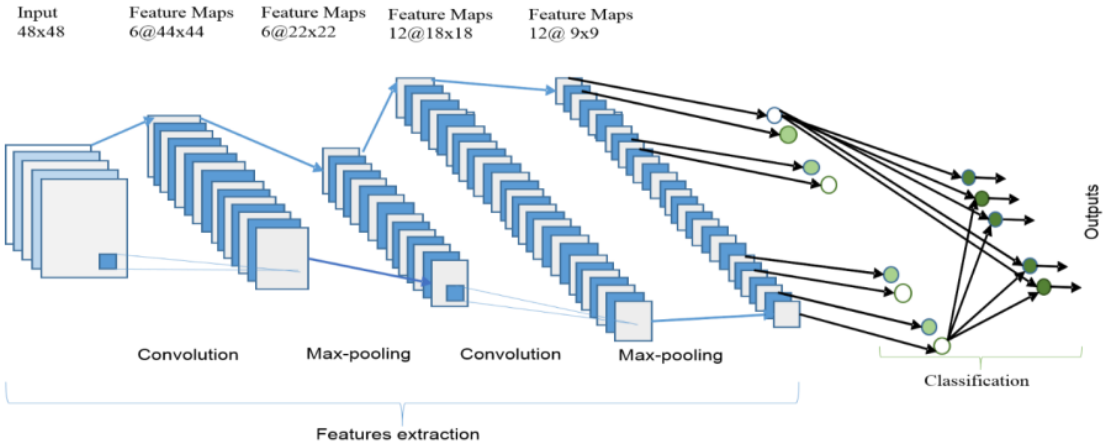


Figure 2.10: The architecture of a CNN composed of alternating convolution and max-pooling layers for feature extraction and a fully connected network for classification [11].

During feature extraction, we call feature maps the output of the convolution and max-pooling and they can be represented on a 2D matrix. As we move through the network, the dimensions of the feature maps decrease due to the pooling operations. Convolution could also reduce the dimensions, but modern CNN architectures use zero padding (adding rows and columns on each side of the feature map to fit the center of the kernel to ensure the same dimension through the convolution operation) to avoid this problem and to be able to apply more layers [12].

Convolution is a type of linear operation where an array, called the kernel, slides through the input and we perform an element-wise product between each element of the kernel and the input. These values are then summed up to obtain the output value on the feature map. The kernel can be seen as a filter containing weights that are learned from the network. These weights are shared across the whole image and are the only learnable components for the convolutional layer. The size of the kernel, stride, padding, and finally, the number of kernels (determining the number of feature maps) are predefined hyperparameters. Figure[2.11] shows an example of convolution using a kernel size of 3x3, no padding, and a stride of 1. It also shows the element-wise product followed by the final output value on the feature map.

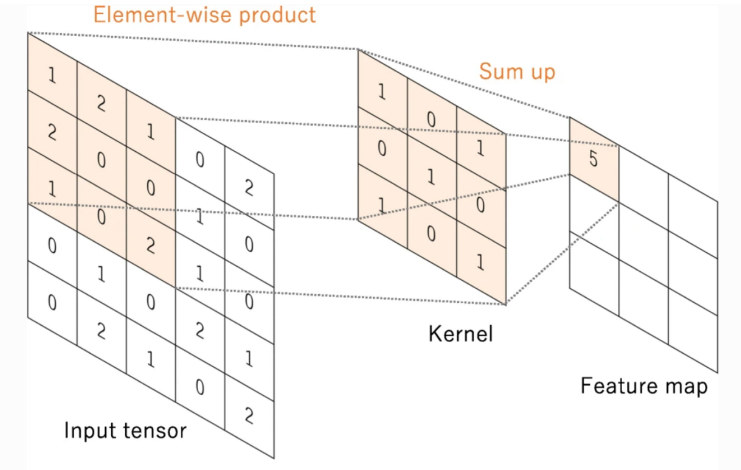
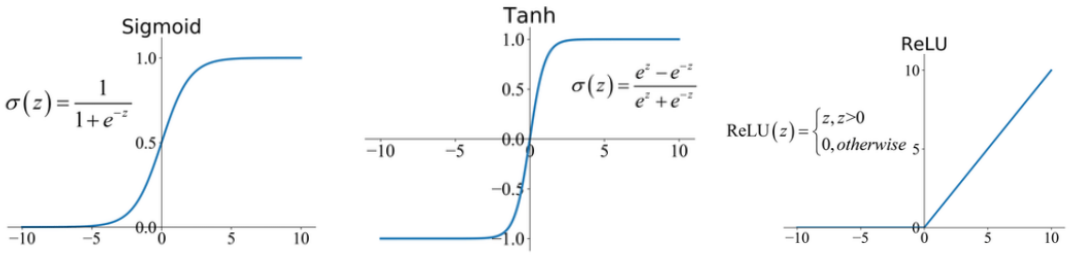


Figure 2.11: Example of convolution using a kernel size of 3x3, no padding and a stride of 1 [12].

The outputs of the convolution are then passed through a nonlinear activation function. The purpose of an activation function is to introduce nonlinearity into the network. It allows for the network to learn patterns that are not linear which would be impossible with a linear (or no) activation function [53]. Sigmoid (Figure[2.12a]) or hyperbolic tangent(tanh : Figure[2.12b]) were originally used as they resemble the most the way neurons operate in our brain [54]. Now, the most successful and widely used activation function is the Rectified linear unit (ReLU) (Figure[2.12c]) [55].



(a) Sigmoid (b) tanh (c) ReLU

Figure 2.12: Visual representation of the 3 most common activation functions [13].

After the convolution, a pooling operation is used to reduce the dimensionality of the feature maps. This operation summarises the feature maps generated by the convolution to reduce the number of parameters to learn for the network. Max pooling is the most common pooling operation, it selects the maximum from a region of the feature map depending on the size of the filter, then the filter slide with a step called "stride" [56]. A common choice is a filter of size 2x2 and a stride of 2 as represented in Figure[2.13], dividing the height and width by a factor of 2. There are no learnable parameters in a pooling operation, the filter size, stride, and padding are hyperparameters predefined, similar to the convolution operation.

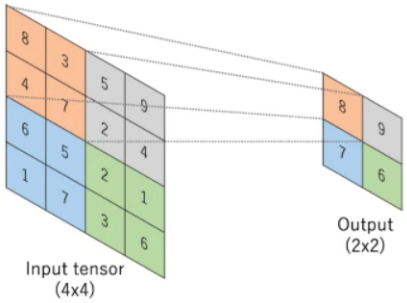


Figure 2.13: Example of a max-pooling operation with a filter size of 2x2, no padding and a stride of 2, dividing the height and width by a factor 2 [12].

The output feature maps of the last convolutional layer are then flattened into a one-dimensional vector and given as input to a fully connected network (composed of fully connected layers) that yields as output the probabilities for each class present in the classification task. This fully connected network is a simple Feedforward Network (FNN), an Artificial Neural Network (ANN) wherein connections between the nodes do not form a cycle [57], the prime example being a Multilayer Perceptron (MLP) which proved to be a universal approximator [58].

The activation function for the last layer of our network (after the last fully connected layer) is usually different from the activation function we presented in Figure[2.12]. Indeed, the output of this activation function will be the output of our network and is, therefore, selected depending on the classification task we perform. For binary classification, a sigmoid is often used and for multiclass classification, we use a softmax function to find the most likely class for a given pixel. This function normalizes the output to obtain values ranging between 0 and 1 and such that the values sum up to 1 and can be interpreted as probabilities.

From CNN to FCN

A Fully convolutional network (FCN) (sometimes called Fully convolutional neural network (FCNN)) is a CNN using only convolutional layers (therefore without the fully connected layers). Instead of using Fully connected layers to produce image classification of the whole image as output, FCNs produce an image of the same size as the input and associate a label with each pixel of the input image obtaining what is also called pixelwise prediction. Figure[2.14] shows the first architecture proposed in 2005 by Long et al. [14].

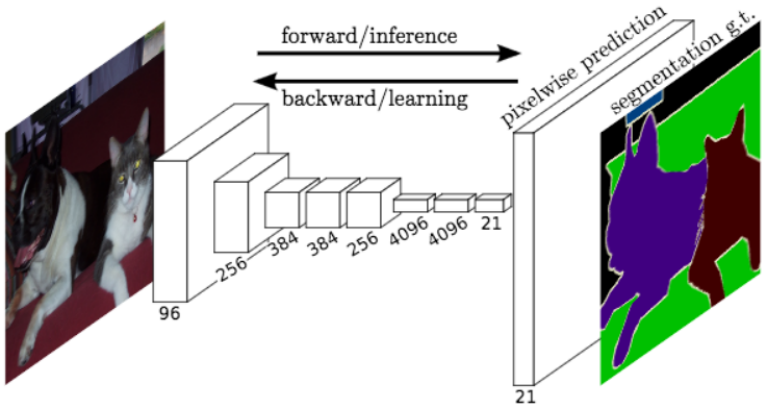
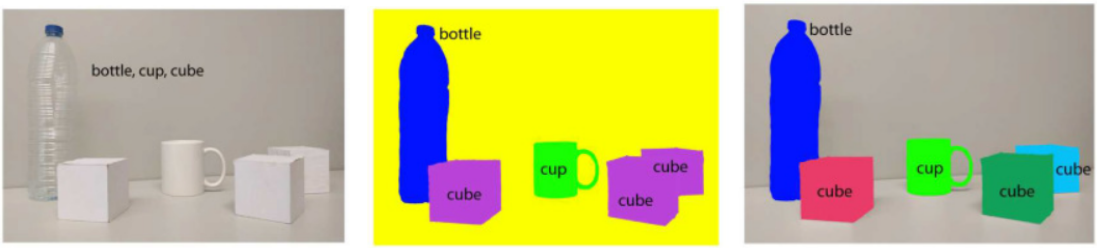


Figure 2.14: FCN architecture proposed by Long et al. in 2005 [14].

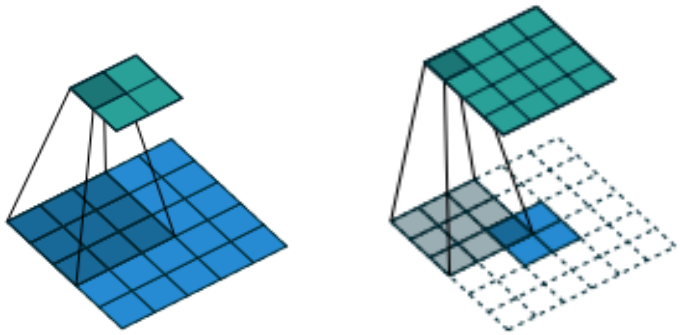
Besides image classification and semantic segmentation, we can also mention instance segmentation. It is a segmentation task where we want to differentiate instances of the same class (different persons on an image, ...), mask R-CNN is an example of a neural network specially built for object instance segmentation [59]. Figure[2.15] gives a visual comparison between image classification that makes a prediction for the whole input, semantic segmentation that makes class prediction for every pixel of the image, and instance segmentation that also gives separate labels for different instances of the same class [15]. For our work, the segmentation of OARs in medical imaging, we are interested in semantic segmentation and this is what we will refer to when we will speak about medical image segmentation.



(a) Image classification (b) Semantic segmentation (c) Instance segmentation

Figure 2.15: Visual comparison between image classification, semantic segmentation and instance segmentation [15].

As we advance through a CNN convolutions and max-pooling layers reduce the dimensions of our image until it reaches the bottleneck. In the second half of a FCN, we have to restore the dimensions of our image. This can be done by upsampling which is also referred to as transposed convolution, up-convolution, or deconvolution. There are a few ways to perform upsampling such as Nearest Neighbor, Bilinear Interpolation, and Transposed Convolution [16]. The last one is the more complex but also the most interesting. It can be thought of as a backward convolution with a kernel that also contains weights to make the upsampling also learnable. Figure[2.16] shows that convolution transpose is essentially the same operation but turned backward.



(a) Regular Convolution (b) Transposed Convolution

Figure 2.16: Visual comparison between regular convolution and transposed convolution [16].

In the same way that transposed convolution acts as reversed convolution, unpooling acts as the reverse of a max-pooling layer. The particularity is that the indices used for the output of the unpooling are remembered from the max-pooling layers as shown in Figure[2.17] [17].

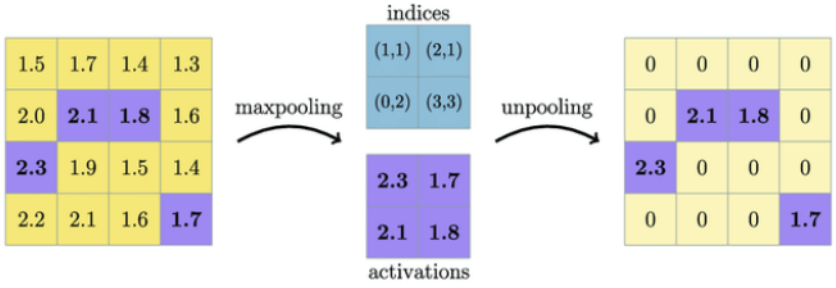


Figure 2.17: Unpooling operation acting as the reverse of the max-pooling operation and remembering the indices used from the max-pooling [17].

An architecture worth mentioning to illustrate the process of upsampling is DeconvNet. DeconvNet uses another network (the Deconvolution Network) on top of the Convolution Network with all convolution and pooling layers reversed. This architecture obtains good resolution on the segmentation but is often criticized for the time it takes to train [18]. Figure[2.18] shows the architecture of DeconvNet.

DeconvNet is not the only neural network developed thanks to the enormous gain in popularity for the use of CNNs and FCNs. The image segmentation literature has lead to a lot of new architectures [60] [61]. Among them, the U-Net architecture stands out as one of the best for medical image segmentation [19] and we will, therefore, discuss it in more detail.

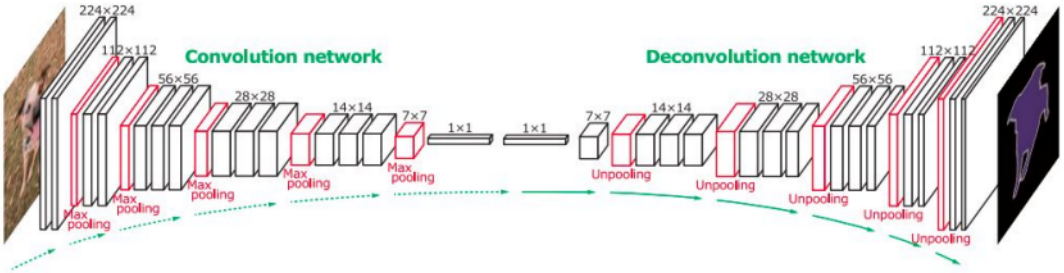


Figure 2.18: The architecture of DeconvNet, showing the downsampling on the left made by a Convolution Network and the upsampling on the right by a Deconvolution network with all convolution and pooling layers reversed [18].

U-Net architecture

U-Net is a model proposed by Ronneberger et al. in 2015, this network is particularly interesting for us because it outperforms other segmentation techniques on a wide variety of segmentation tasks [19]. It gets its name from its "U" shape. The global architecture is composed of a contracting path to capture context and an expanding path that enables precise localization. Figure[2.19] shows the global architecture of U-Net.

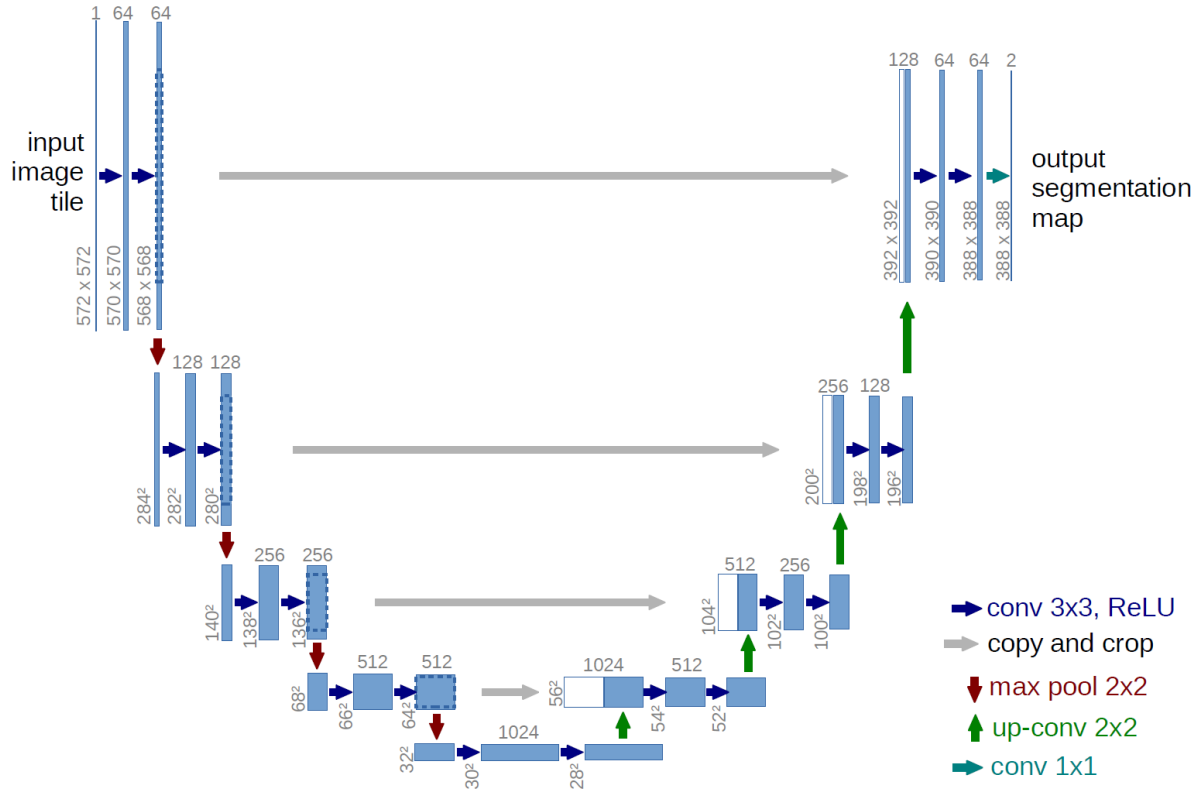


Figure 2.19: The U-Net architecture consisting of a contracting path to capture context and an expanding path that enables precise localization [19].

The contracting path consists of a succession of 3x3 convolution with ReLU activation function, followed by a 2x2 max-pooling. After the contracting path, the network reaches the bottleneck that contains a very compressed representation of the data to help the model extract high-level features. The expanding path consists of a sequence of 2x2 up-convolution and concatenation with the features maps copied (the white boxes on Figure[2.19]) from the contracting path through what is called a "skip". This architecture allows U-Net to learn features in multiple resolutions. The U-Net architecture can, more intuitively, be thought of as an encoder-decoder model with the contracting path being the encoder network and the contracting path being the decoder network.

Many models use the U-Net architecture as a basis to come up with other architectures. For example, Deep Residual U-Net combines the strengths of U-Net and residual learning. It has been proposed successfully for Road Extraction from aerial images obtaining great success [62].

Problems with CNNs in medical image segmentation

Despite their great performance in the field of medical image segmentation, the use of CNN also comes with some problems. One of the biggest concerns with the use of CNNs in medical image segmentation is that they require a sufficiently large and diverse annotated dataset. This is often hard to gather due to the privacy regulations [63] [64] and the time and high level of expertise required to annotate the data.

Another big concern with the use of CNNs in medical image segmentation is their lack of interpretability. Indeed, by their nature of "deep" neural networks, CNNs are very complex and it's almost impossible to trace back what made the model classify a pixel with a certain label. The complexity of the models and the large number of hidden layers make it difficult for a human to understand how they reach their conclusion, which leads to their use as a "blackbox" [29].

2.2 Active learning

In medical image segmentation, the best performing models are often based on deep neural networks, but these networks require large and diverse datasets. These are often unavailable due to the privacy regulations [63] [64] and the time and high level of expertise required to annotate the data. This shortage is even more present for annotated CBCT scans, which we need to monitor anatomical deformations between planning and treatment sessions in radiotherapy. This situation is a real problem for the deployment of deep neural networks in the medical field.

Active learning addresses this problem by choosing to give the model data that will improve the model the most, and asking the expert (called the oracle) to annotate this particular data. The goal of active learning is to achieve better performance with fewer annotated data by allowing the model to choose the data from which it learns [20]. Active learning integrates the oracle inside the learning process by sending him queries, these queries take the form of unlabeled data that we ask the oracle to label. Figure[2.20] represents the three scenarios possible depending on how our active learning model creates the queries it will send to the oracle.

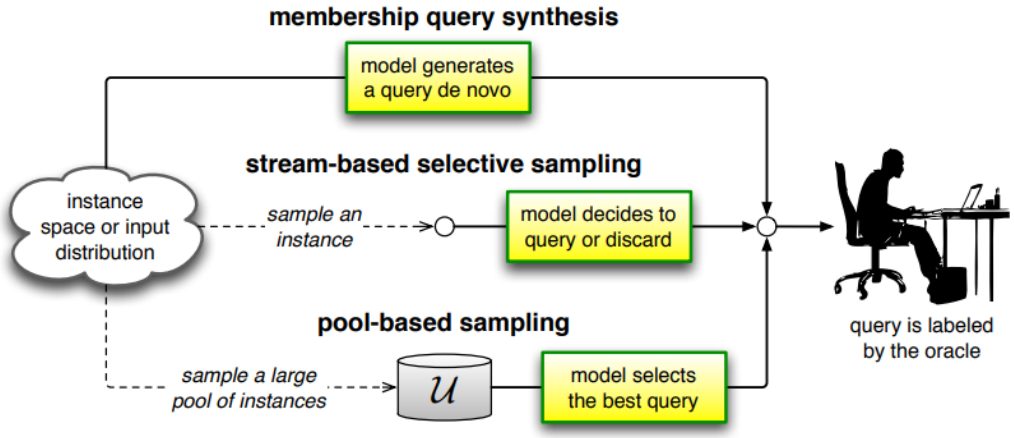


Figure 2.20: The three scenarios of active learning [20].

- **Membership query synthesis** : The model generates new unlabeled instances and asks the oracle to label them. The limitation of this scenario is that, sometimes, the model can create unlabeled instances that are not recognizable by a human annotator as happened in [65] for the classification of handwritten characters.
- **Stream-based selective sampling** : The unlabeled data are sent, as a stream, one by one to the model. The model can either decide to discard the instance or send a query to the oracle to label it [66].
- **Pool-based sampling** : The model selects the best instance (or batches) to query from a pool of unlabeled instances based on a given sampling strategy [67].

Methods based on pool-based sampling can be computationally expensive, as each iteration requires an evaluation of the entire unlabeled data pool. However, these methods have shown the most promising results when combined with Deep Learning methods [68] and are, therefore, the most studied in the literature. The big advantage for us is that it makes its decision based on

Pool-based sampling is an iterative process repeating itself until a certain budget is exhausted, a certain level of performance is reached or a stopping criterion is met [69]. The pool-based sampling cycle is summarised in Figure[2.21]. the model is trained with a certain training set. Then, it selects from the unlabeled pool the queries it wants the oracle to label. Finally, the oracle label the data expanding the labeled training set, and the loop can continue.

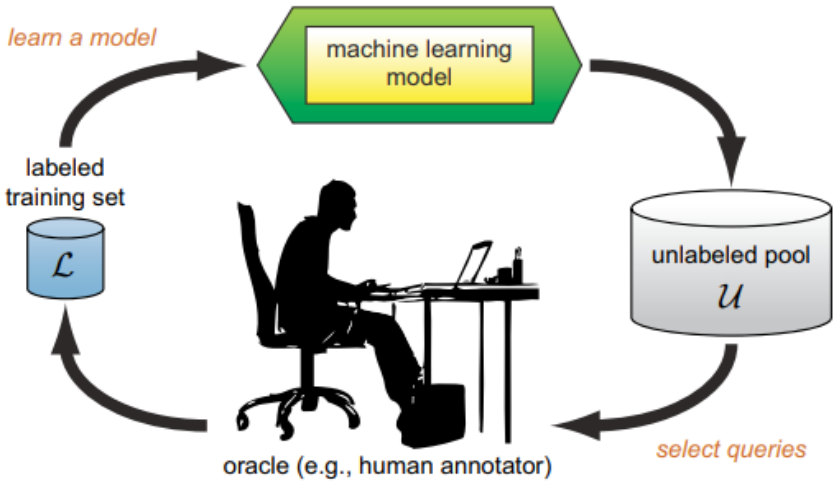


Figure 2.21: The pool-based sampling cycle [20].

The most crucial part of the pool-based sampling cycle is query selection. To perform its selection, the model has to define what it is looking for. We can identify two broad categories of query strategies: informative-based query strategies, which decide the queries based on the informativeness of an instance, and representative-based query strategies, which select queries based on the representativeness of an instance [70].

Informative-based query strategies are the most common. They are based on the evaluation of the informativeness of the instances. The more informative an instance is, the more likely it is to be queried. Two of the best-known query strategies based on the measure of informativeness are uncertainty sampling and query by committee. In uncertainty sampling, we ask the oracle to label the instances for which the model is least certain. In query-by-committee, we consider the opinion of a committee of models and ask the oracle to label the instance they disagree the most on, this can sometimes be seen as a certain type of uncertainty sampling where the committee is used to assess uncertainty. We will have the opportunity to discuss uncertainty sampling and query-by-committee in more detail later.

On the contrary, representative-based query strategies are based on evaluating the representativeness of the instances. These models use the structure of the unlabeled pool of data to measure how well the selected instances represent the overall input patterns of the unlabeled data [71]. It can often be interesting to add a measure of representativeness to a model already using a measure of informativeness. For example, when multiple instances are chosen based on the same informative criteria it may lead to redundancy in the labeled set [70]. The diversity-based approach can be used to deal with this problem by trying to preserve diversity in the dataset. This approach was investigated by Hoi et al. for medical image classification [72]. On the other hand, if we rely only on diversity, the model can choose data with which it will not learn anything.

There are other types of query strategies, such as expected error reduction, which optimizes the impact of adding a certain instance on the expected future error of the model [73]. Or expected gradient length, that chooses the instance that would cause the greatest change in the model if we knew its label [74]. We analyze deeper uncertainty sampling and query-by-committee because they are the most common in the literature and our approach follows the query-by-committee strategy.

Uncertainty Sampling

Uncertainty sampling was proposed by Lewis and Gale in 1994 [67]. The key assumption in this query strategy is that the more uncertain a prediction is, the more information we can gain by labeling that instance and including it in our training set [68].

For multi-label classification, several measures of uncertainty are commonly used, either selecting the data for which the model is the least confident, or taking the first and the second most probable label into account (called margin sampling) [75], or using entropy as a measure of uncertainty [76].

Many active learning models are built around this idea of measuring uncertainty, and many applications were born from this idea. For example, uncertainty sampling was applied to support vector machines for text classification [77]. Another example is Bayesian neural networks that integrate dropout inside a neural network to measure uncertainty in deep neural networks [78].

Query-By-Committee

The query-by-committee algorithm proposed by Seung et al. chooses the next query according to the principle of maximal disagreement [79]. The fundamental principle of the query-by-committee approach is to measure informativeness by looking at the disagreement between the members of the committee. In this approach, we consider that the higher the disagreement on a given instance, the more information this instance will give to the model. There is no consensus on the optimal size of the committee, most of the time the size used varies depending on the application and model architectures. Even models using a committee of 2 members have shown great results [80].

Ensemble methods can be seen as methods based on the query-by-committee approach. Ensemble methods train multiple models, often with the same architecture under different initialization or hyperparameters, and use the output of all these models as a way to measure uncertainty [81].

Recently, Beluch et al. [81] have shown that ensembles obtain great performance in active learning outperforming other methods of uncertainty estimation such as Monte-Carlo Dropout. They explain this by the fact that their method creates models with higher diversity than models created with Monte-Carlo Dropout.

Ensemble methods perform well to measure informativeness but they have a high computational cost because they need to train multiple models and each model needs to be updated when new training samples are added [68].

2.3 Performances Measurements

For semantic segmentation, we assign each pixel in the image a label. Labeling each pixel of the image can be seen as a classification problem for each pixel. When we compare the prediction to the ground truth, for each pixel we have four possible outcomes, the pixel is either a True Positive (TP), a True Negative (TN), a False Positive (FP), or a False Negative (FN). Figure[2.22] summarize these different results in what is called the confusion matrix.

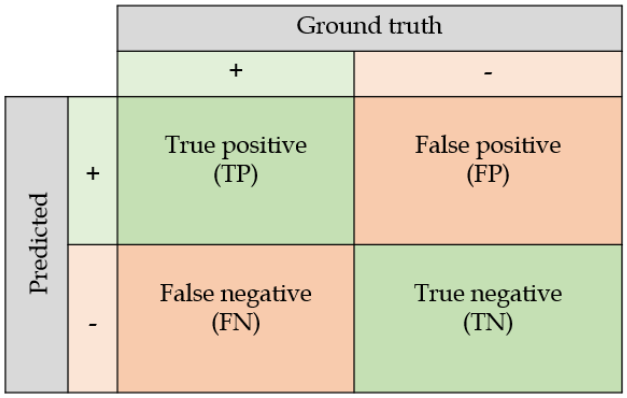
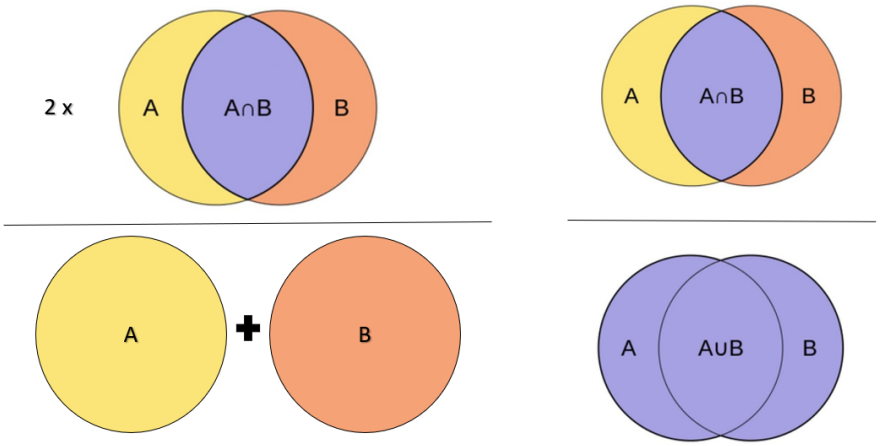


Figure 2.22: Confusion Matrix.

For pixel-wise segmentation, we want to measure the performance of our prediction by comparing the overlap between the prediction and the ground truth. Two metrics based on the measure of the overlap are often used Intersection over Union (IoU) (also known as the Jaccard Index (JI)), and the Dice similarity coeficient (DSC) (often referred to as Dice).

$$Dice = \frac{2|A \cap B|}{|A| + |B|} = \frac{2TP}{2TP + FP + FN} \qquad IoU = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN}$$

Figure[2.23] shows how these metrics can be interpreted as a measure of the overlap between the prediction and the ground truth.



(a) Dice coefficient (b) Intersection over Union

Figure 2.23: Visualization of IoU and Dice coefficient to measure overlap.

Both the Dice coefficient and the Jaccard index have become some of the most commonly used metrics for the evaluation of segmentation tasks in medical imaging [82]. We could have chosen either of them but decided to stay with the Dice coefficient as it was the performance measurement we encountered the most in the literature.

Chapter 3

Proposed approach

Deep learning models have shown great performance applied to the medical field. For example, the U-Net architecture outperforms other segmentation techniques for a wide variety of segmentation tasks in medical imaging [19]. However, they require a sufficiently large and diverse annotated dataset. With CBCT images, this is often unavailable for various reasons:

- Manually annotating CBCT images requires a high level of expertise and is very time-consuming for the radiologist.
- The European legislation [63] surrounding the preservation and protection of privacy for medical information makes it harder to gather the data necessary [64].
- Models trained with CT data do not generalize well to CBCT images [83]. Therefore, we have to find CBCT images manually segmented but they are not annotated during the typical radiotherapy workflow.

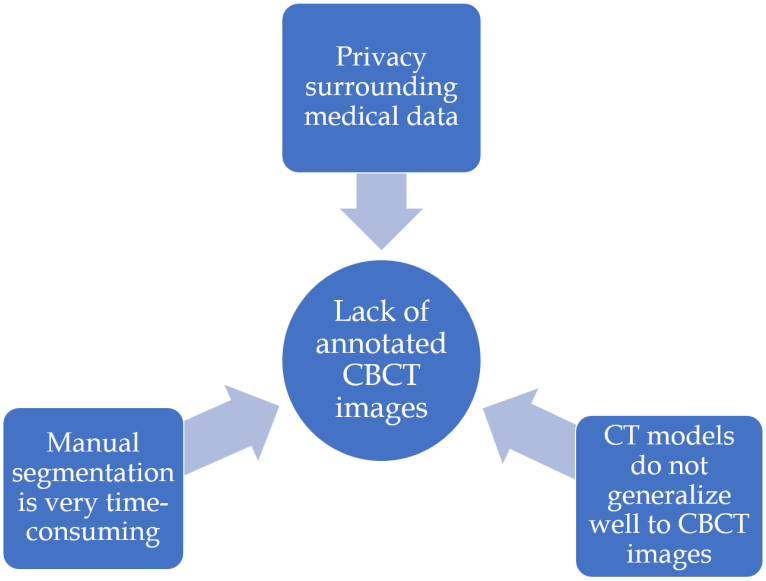


Figure 3.1: Reasons for the lack of annotated medical data with a focus on CBCT images.

In this section, we want to address this problem and we propose to reduce the amount of data needed for training by specifically labeling the data that provides the most information. We propose an active learning model, following the query-by-committee approach. Our model uses the disagreement between a 3D fully convolutional U-Net and a diffeomorphic Morphon algorithm to make an informed decision on the best data to add to the training dataset. Liu et al. [84] have shown that anatomical information can be incorporated as an enhancement in many steps of DL model deployment. We have integrated this idea into our model by using atlas-based segmentation (with the Morphons algorithm) to obtain a second opinion containing anatomical information.

Recently, ensemble techniques have been very successful in active learning [81]. However, these techniques often rely on comparing the opinion of several models following the same architecture trained with different hyperparameters and initialization. On the contrary, our approach is based on the idea that active learning would benefit most from comparing the opinion of models with completely different architecture. Comparing the opinion of a deep learning model to an algorithm based on atlas segmentation is a good way to obtain two completely different opinions based on two completely different architectures and make the most informed decision.

For our work, we used CT and CBCT images of the pelvis in prostate cancer. The CT scans are taken and annotated during the planning of the radiotherapy treatment session. However, between planning and the day of treatment, large deformations may occur and lead to a higher dose being delivered to the surrounding healthy organs [3]. To monitor deformations, CBCT scans are taken on the day of treatment. The segmentation task is to contour the bladder and the rectum on the CBCT images. The 117 data were divided into 3 folds, 2 folds for training and 1 fold for testing.

Our active learning approach is an iterative process. We start with a small training dataset of 5 data and, at each iteration, we select from a pool of unlabeled data 5 instances that are annotated and added to the training dataset. The Morphons algorithm performs image registration on the CBCT image using the CT image of the same patient. Therefore, its prediction is independent of the number of training data available and can be computed once at the beginning of the program. On the contrary, the U-Net algorithm is trained using annotated CBCT images. It can achieve higher accuracy than the Morphons algorithm but is dependent on the number of images it can use for training. It must be trained again each time the training dataset is increased.

Figure[3.2] represents the global structure of the program that can be described in a few steps :

1. The predictions of the Morphons algorithm are computed at the start of the program.
2. The U-Net model is trained with the training dataset.
3. The predictions of the U-Net model are computed.
4. We select the data based on the disagreement between both predictions.
5. The data selected are added to the training set and the loop goes back to step 2.

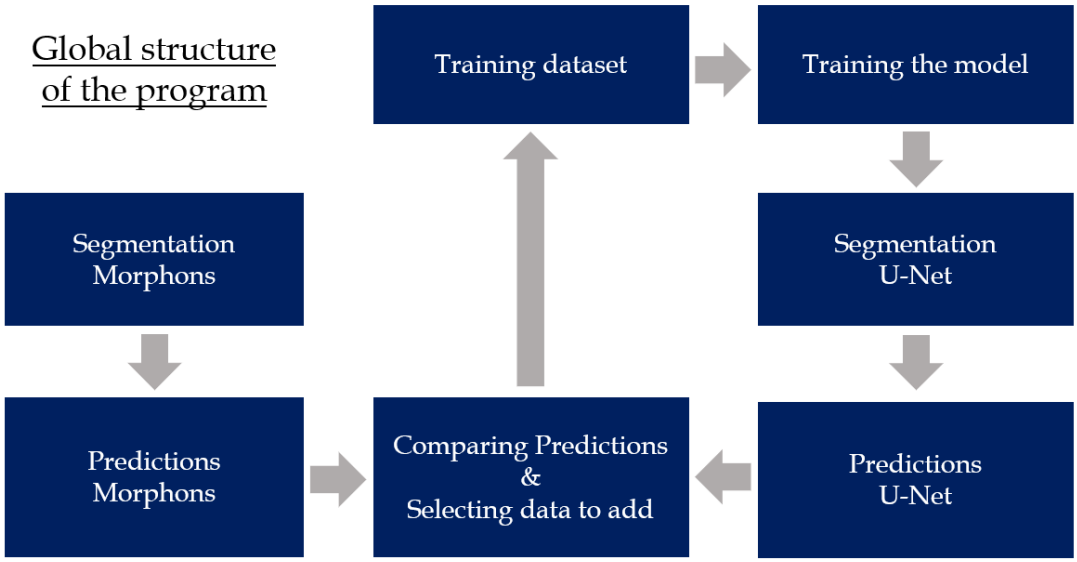


Figure 3.2: Global structure of the program.

Our model uses the Dice score between the two predictions as a measure of disagreement. A low dice score means that the two models disagree a lot and the instance will be more likely to be queried. Our model selects the 5 data with the lowest average dice score, averaging between the Dice obtained for the bladder and the Dice obtained for the rectum. Figure[3.3] shows the 5 data selected on the iteration going from 60 to 65 training data according to the dice value between the U-Net prediction and the Morphons prediction.

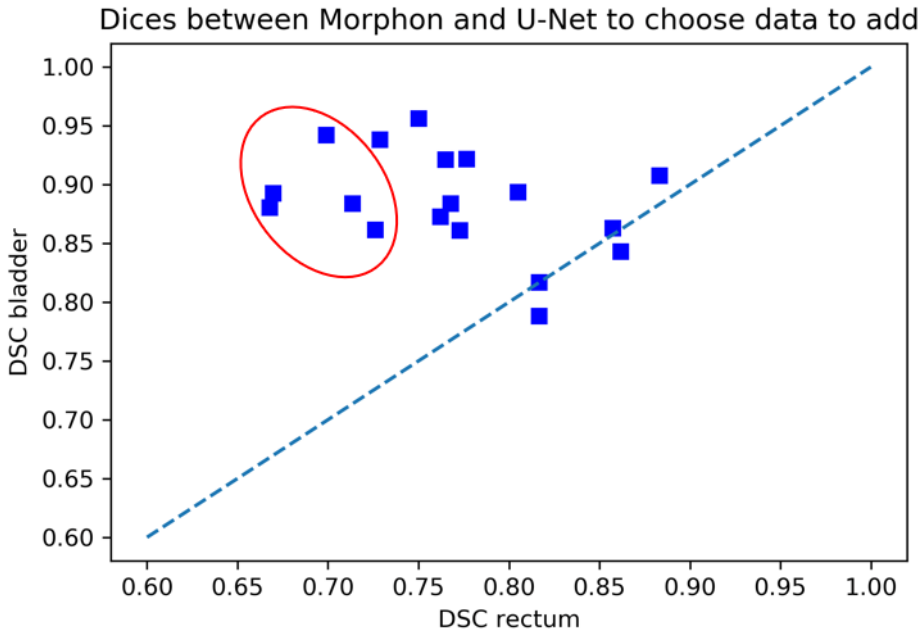


Figure 3.3: Visualization of the data selection.

Chapter 4

Experimentation

This chapter is divided into 3 parts. We begin by discussing the experimental methodology we followed. We explain the dataset, show a summary of the dataset at our disposal, and give a visualization of the data. We also explain the implementation of the algorithms, the preprocessing, and the hyperparameters tuning. Next, we present our results that tested our approach against random selection for bladder and rectum segmentation. Finally, we conclude with a discussion of our results and the limitations of the approach.

4.1 Experimental methodology

Dataset

The data used for this work came from CHU-Charleroi Hôpital André Vésale and CHU-UCL-Namur and were delineated by a trained expert for the study in Léger et al. [5].[1] The initial dataset was a combination of CT and CBCT scans of the pelvis, but not all organs were manually annotated on each scan. Table[4.1] shows a summary of the data we had at our disposal.

hospital & scans	bladder	rectum	prostate
Charleroi CT	88	80	75
Charleroi CBCT	73	85	22
Namur CT	138	139	138
Namur CBCT	86	86	41

Table 4.1: Summary of the data at our disposal.

Our goal is to perform a segmentation of the organs at risk (the bladder and the rectum) to ensure their radiation exposure remains within safe limits. We, therefore, selected patients on whom we had both a CT and a CBCT scan, and their annotated version containing the bladder and the rectum. This left us with 43 patients from Charleroi and 74 from Namur, for a total of 117 patients.

Data visualization

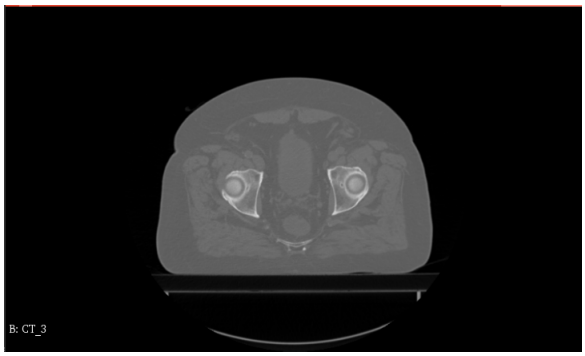
Figure [4.1] gives a visualization of the CT and CBCT scans that we work on with 3 different views: axial, coronal, and sagittal.

Figure[4.2a] and Figure[4.2b] give a 3D visualization of the annotated version of both scans.

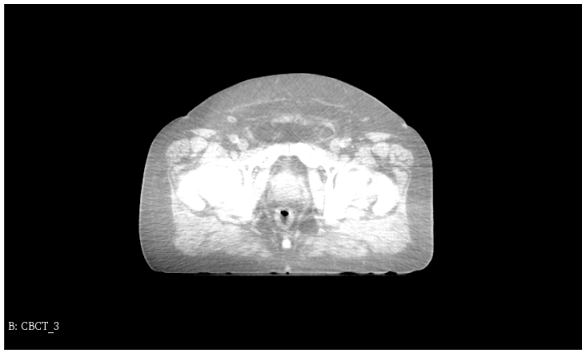
Preprocessing

For Morphons registration, we didn't crop the data but for U-Net segmentation, all the images and masks were cropped to volumes of 128 x 256 x 128 to reduce memory consumption. We used standardization, rescaling the image intensities so that they are centered around 0 with unit variance, because neural networks trained with standardized data yield better results, especially when the number of data available is small [85]. Our U-Net architecture uses data augmentation because it improves diversity and decreases the risk of overfitting [86]. The training set is augmented using rotation (-5° to +5° along each of the three axes), shift (-5 to +5 pixels along each axes) and shear [5].

[1]Special thanks to Jean Léger and Eliott Brion for the opportunity to work on real data.

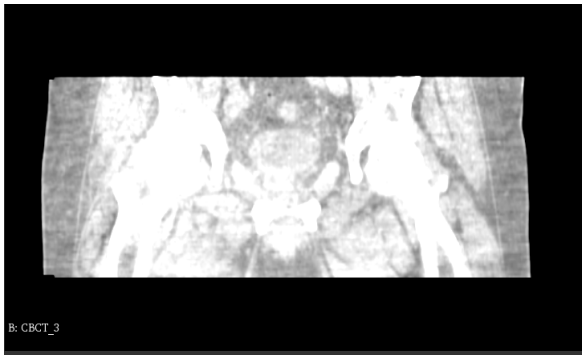


(a) CT axial view

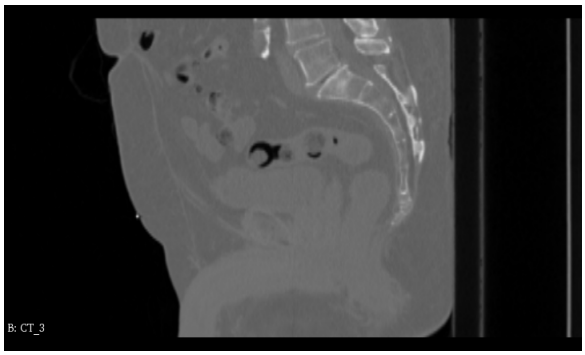


(b) CBCT axial view

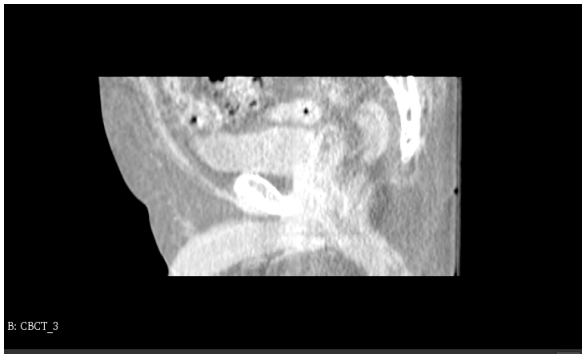
(c) CT coronal view



(d) CBCT coronal view

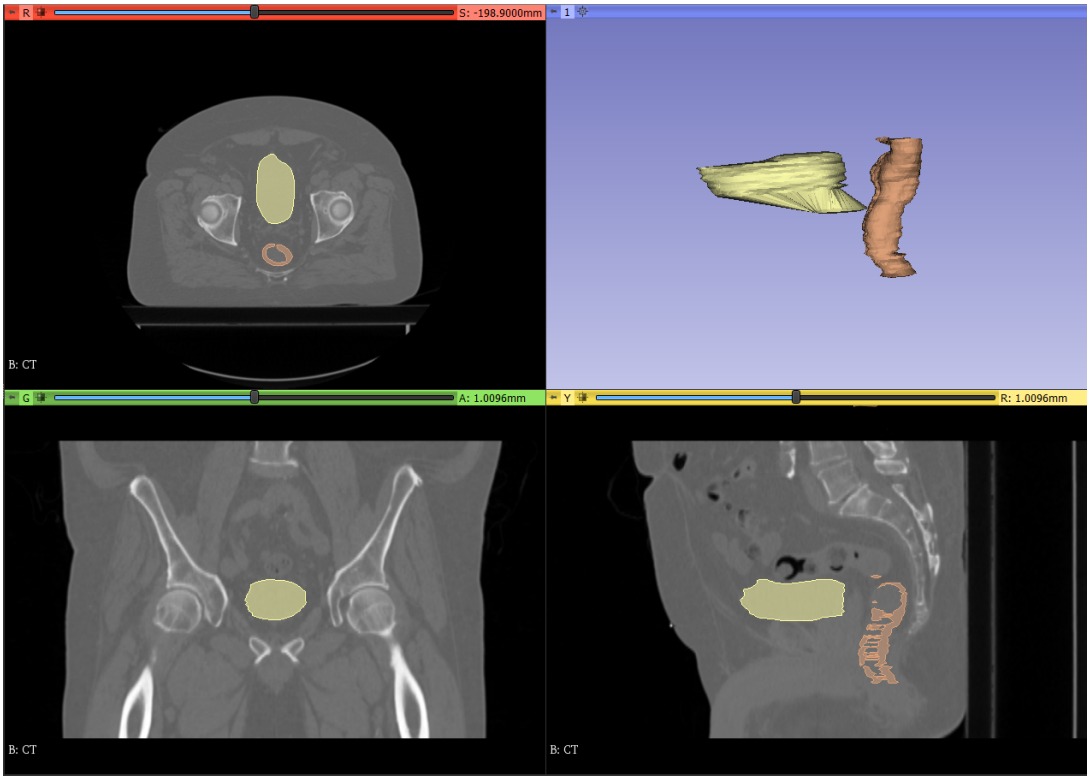


(e) CT sagittal view

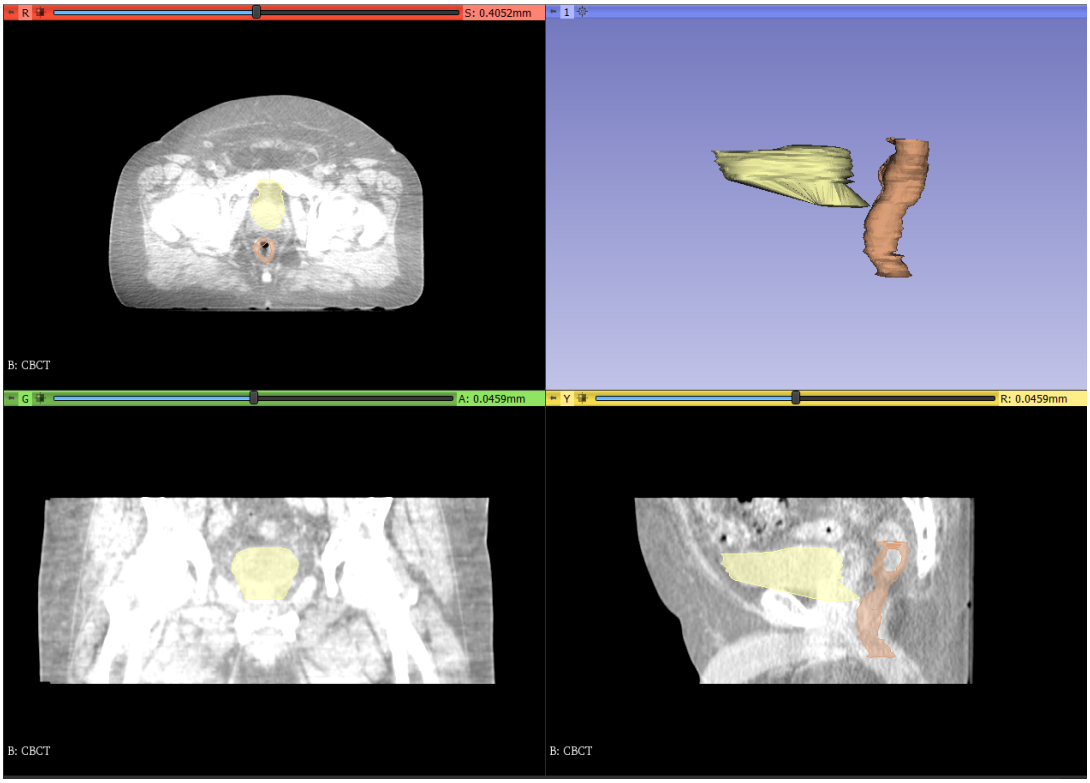


(f) CBCT sagittal view

Figure 4.1: Visualization of CT and CBCT scans in axial, coronal and sagittal views.



(a) Annotated CT scan



(b) Annotated CBCT scan

Figure 4.2: Visualization of the bladder and rectum on annotated CT and CBCT scans.

Implementation of the algorithms

For the U-Net implementation, we used a 3D U-Net fully convolutional neural network based on the python implementation developed by Brion et al. for the study in [87] and [5]. The implementation was made publicly available at `https://github.com/eliottbrion/pelvis_segmentation`. We chose to use the U-Net algorithm because of its high performance on segmentation tasks [19].

For the Morphons implementation, we used a diffeomorphic version of the Morphons algorithm implemented in OpenReggui (`https://openreggui.org/`) [41]. We chose the diffeomorphic Morphons algorithm for registration because it is suited for registering planning and daily images and it enforces deformations that are physically possible [41].

Hyperparameters tuning

Our U-Net implementation uses the Adaptive moment estimation (Adam) optimization algorithm, an optimization algorithm for gradient descent [88]. The learning rate controls how quickly the model is changing [89], we used 10^{-4} because a higher learning rate can make the network unstable [90] but a smaller learning rate increases too much the computational time. We used 100 epochs because it was enough for our model to converge in all the configurations used. We had to increase the depth of our model from 6 to 8 layers for the two organs to converge. Finally, we reduced the batch size from 2 to 1 to meet memory constraints.

For the loss of our model, we compared different possibilities (for example the categorical cross-entropy) and stayed with a Dice loss (an average of the dice obtained on the rectum and the bladder) because it obtained the best results. Table [4.2] gives a summary of the hyperparameters used.

Hyperparameter	Value
loss	Dice loss
optimizer	Adam
learning rate	10^{-4}
epochs	100
layers	8
batch size	1

Table 4.2: Hyperparameters of the U-Net algorithm.

4.2 Results

We tested our approach against random selection for bladder and rectum segmentation. Performance is measured by computing the Dice score between the prediction and the ground truth. Figure[4.3] shows the performance measurements for bladder segmentation (on the left) and rectum segmentation (on the right) as a function of the number of training data available.[2]

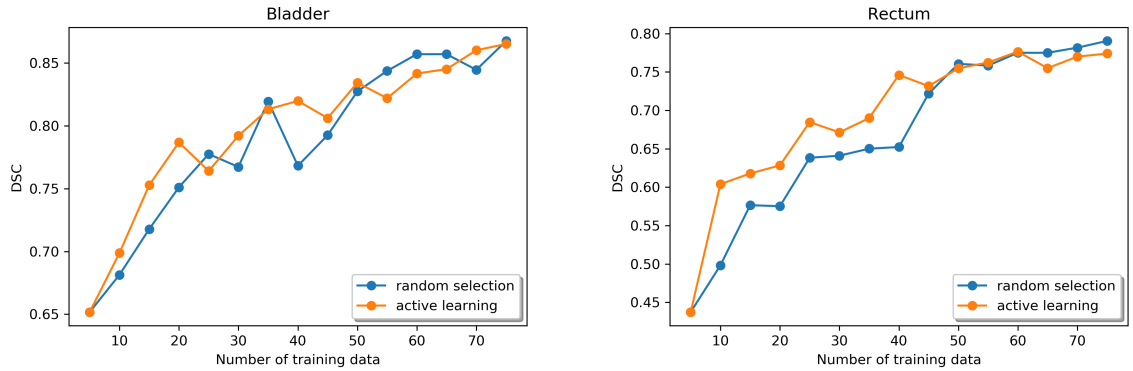


Figure 4.3: Performance of our active learning approach compared to random selection.

Figure[4.4] shows the loss of our model compared to random selection. The loss of our model is a dice loss calculated as an average between the dice values of the two organs. It may be more informative to look directly at the dice loss as it represents the performance of the model for the whole segmentation.

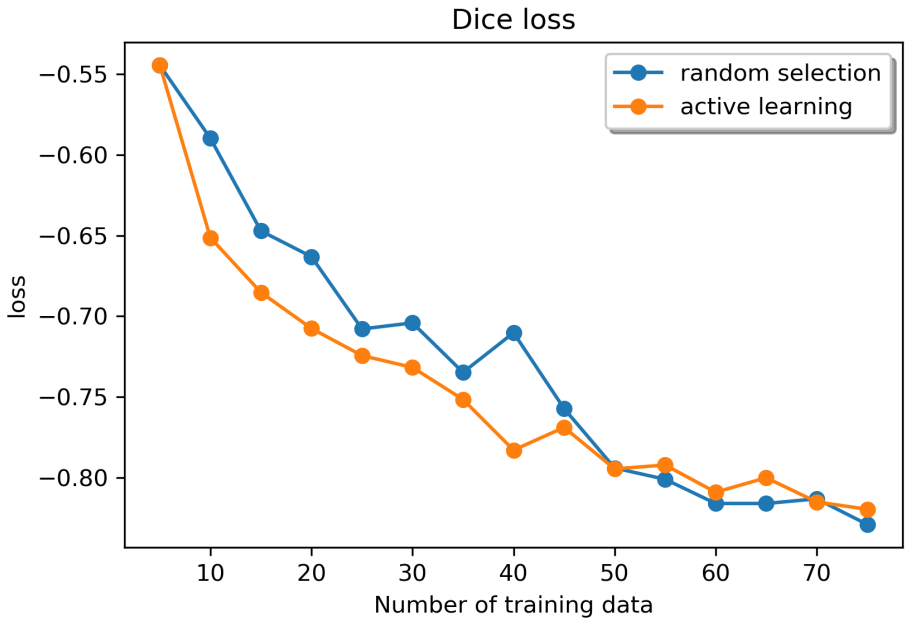


Figure 4.4: Dice loss for our active learning approach compared to random selection.

[2]All computations were performed on a server at UCLouvain.

4.3 Discussion

Figure[4.5] shows that, for rectum segmentation, our active learning approach clearly outperforms random selection when the model is trained with less than half of the available dataset. For bladder segmentation, the results are less clear, our active learning approach tends to perform better on small training datasets but the results are noisy. The results also show that, on our dataset, rectum segmentation was more challenging for the model than bladder segmentation and we obtain lower dice values for rectum segmentation across the board. By looking at each organ separately, we can ensure that our approach performed well on each organ and did not focus on one specific organ at the expense of the other's performance.

Instead of looking at each organ separately, Figure[4.6] shows the performance of our approach for the whole segmentation. This is often more informative because the U-Net model we used optimizes the dice score for the entire segmentation. Figure[4.6] shows convincing results, it proves that data selection with our active learning approach allows the U-Net model to reach higher performance than random data selection. This demonstrates that we can compare the predictions of atlas-based segmentation to the predictions of a deep learning model and use this comparison to intelligently query the oracle to reduce the amount of data necessary for training.

Our approach is especially effective at the early stage of the process. With a training set larger than 40, which is half of the available dataset, the two methods are relatively equal. This is to be expected because the advantage of selecting the data is very significant at the beginning but becomes less significant as the number of training data increases. For example, at the end of the selection, the two approaches have selected all the data available and end up with the same training set.

Limitations of the approach

We compared our approach to random selection, which shows that our method works, but we didn't compare it to other active learning strategies. It might be interesting to implement a state-of-the-art ensemble technique using different models with the same architecture to see if using models with completely different architectures was beneficial.

Our approach makes its decision by comparing the prediction of two models. It could be interesting to integrate into the approach more models following different architectures. For example, we could also use the opinion of a model based on active contour. At the moment, our approach gives the same importance to each opinion. If we integrate more models, it could be interesting to give a weight to each opinion, representing their importance, and find an optimal combination.

Our approach assesses the informativeness of an instance under the assumption that the more the two models disagree on an instance, the more informative it would be for the model to add that instance. Like other information-based active learning strategies, when many instances are selected at once it can lead to a decrease in the representativeness of the training dataset. To deal with this problem, some strategies also integrate a measure of representativeness such as a diversity-based approach.

Chapter 5

Future work

5.1 Second opinion for a safer deployment

In the last chapter, we described an active learning approach that reduces the amount of data needed to train a deep learning model. In this chapter, we explore the idea of relying on a second opinion when the amount of data is low to maintain good performance. This would allow for a safer deployment of the deep learning model because it can use the second opinion when the training set is too small.

5.1.1 Description of the approach

We propose an intelligent system deciding between both models depending on the number of labeled data available for training. If the number of annotated data is sufficient, our system uses the prediction of a U-Net model for its high performance on most segmentation tasks [19]. When the number of annotated data available for training is too small, it uses atlas-based segmentation because this technique relies on image registration and, therefore, doesn't require training [36]. Figure[5.1] illustrates the two scenarios.

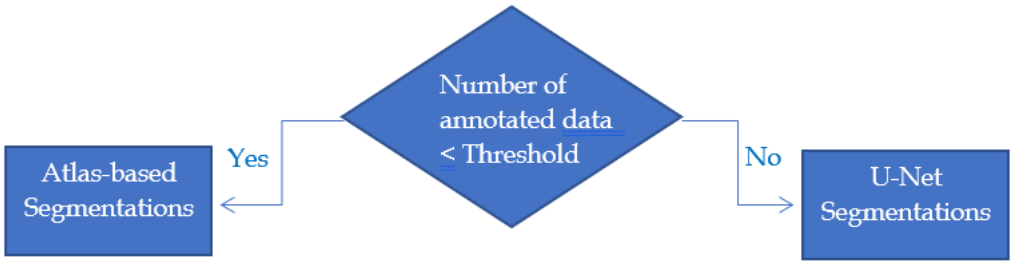


Figure 5.1: The two scenarios of our mixed approach.

Our mixed approach uses a diffeomorphic Morphons algorithm for atlas-based segmentation because it enforces deformations that are physically possible [41], its implementation is available at `https://openreggui.org/`. For the U-Net segmentation, it uses a fully convolutional 3D U-Net developped by Brion et al. for the study in [87] publicly available at `https://github.com/eliottbrion/pelvis_segmentation`.

5.1.2 Results

We tested our mixed approach for bladder and rectum segmentation and compared it to U-Net segmentation and Morphons registration. Figure[5.2] shows the performance measurements for bladder segmentation (on the left) and rectum segmentation (on the right) as a function of the number of training data available. Both performances are evaluated by computing the Dice score between the prediction and the ground-truth.

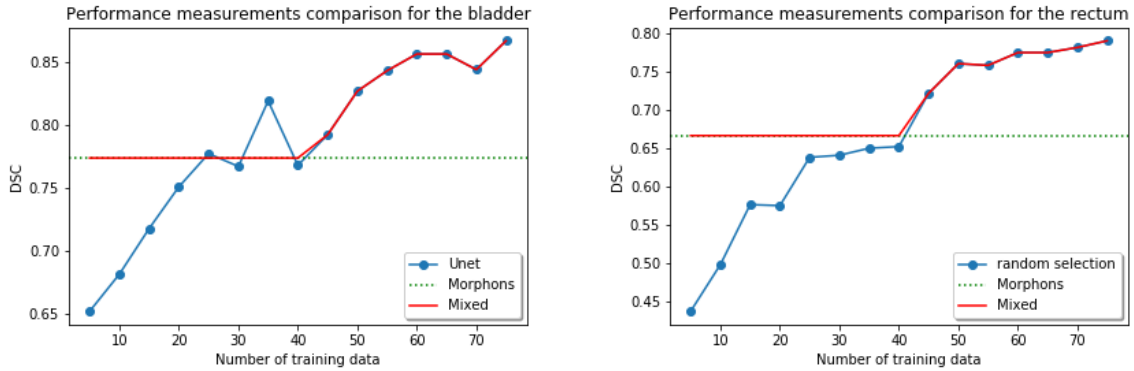


Figure 5.2: Performance of our mixed approach compared to U-Net segmentation and Morphons registration for bladder segmentation (left) and rectum segmentation (right).

Figure[5.3] shows the performance on the whole segmentation by averaging the dice value we obtained on the two organs. This can be more telling as our algorithms are optimized to obtain the best performance on the whole segmentation.

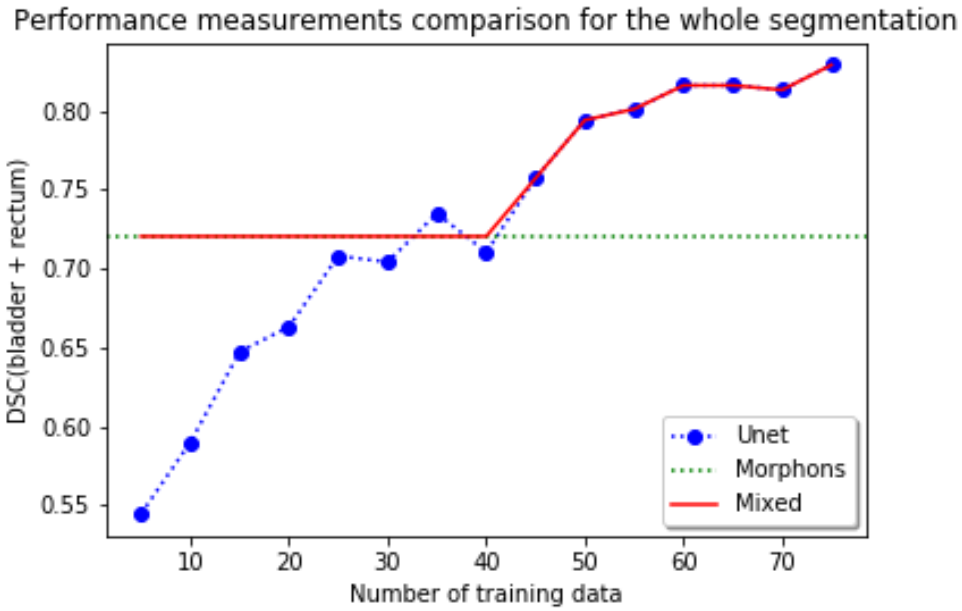


Figure 5.3: Performance Measurements of our mixed approach compared to U-Net segmentation and Morphons registration for the whole segmentation.

5.1.3 Discussion

Figures[5.2] and [5.3] show that the Morphons algorithm outperforms the U-Net implementation when the training set is small. On the contrary, the U-Net implementation outperforms the Morphons algorithm when the size of the training set increases. These results are consistent with those obtained in Léger et al. [5] that compared a U-Net implementation to a Morphons algorithm and showed that with a lack of data for its training, the U-Net model underperforms the Morphons algorithm. Our mixed approach manages to outperform both techniques because it takes the best of both models. When the training dataset is too limited, it relies on atlas-based segmentation and when the training dataset is large enough it switches to U-Net segmentation to reach the best performance possible.

The mixed approach we propose still needs more work for a real-world implementation. The method switches from one segmentation technique to the other if the number of annotated data is bigger than a certain threshold (Figure[5.1]). But we didn't discuss a way to define this threshold. For our computation, we used 40 training data because it was the crossing point between U-Net and Morphons. In a real-world implementation, we wouldn't know this in advance and we would need a way to decide when to switch from one to the other.

Even though the method would need more work, it shows the relevance of relying on atlas-based segmentation when large and diverse datasets are unavailable. This perspective can also be interesting during the deployment of a deep learning model when all the necessary data are not yet gathered. Atlas-based segmentation can, then, be used as a robust second opinion to obtain safer predictions.

Chapter 6

Conclusion

This master thesis proposed an active learning approach to reduce the amount of labeled data needed to train a deep learning model. It is particularly useful in medical image segmentation, as the best performances are achieved by deep neural networks but these networks require large and diverse datasets. This is often hard to gather due to the privacy regulations and the time and high level of expertise required to annotate the data. The proposed approach follows a query-by-committee strategy that compares the prediction of a U-Net model to the prediction of a Morphons algorithm. Their disagreement is then used to specifically label the most informative images to add to the training set. We tested our approach against random selection for bladder and rectum segmentation.

The results showed that the U-Net model trained on data selected with our approach outperformed the U-Net model trained with data randomly selected. The advantage of choosing the data was most significant at the start of the process but became less meaningful when the number of training data increased. Nonetheless, it demonstrated that we can compare the predictions of atlas-based segmentation to the predictions of a deep learning model and use this comparison to intelligently query the oracle to reduce the amount of data necessary for training.

This master thesis also proposed an intelligent system, using a U-Net model when the amount of training data is sufficient and relying on a Morphons algorithm otherwise. The results proved the relevance of relying on atlas-based segmentation when large and diverse datasets are unavailable and showed that atlas-based segmentation can be used as a robust second opinion during the deployment of a deep learning model.

Bibliography

[1] Huynh E., Hosny A., Guthier C., Bitterman Danielle S., Petit Steven F., Haas-Kogan Daphne A., Kann Benjamin, Aerts H. J. W. L., and Mak R. H. Artificial intelligence in radiation oncology. 17(12):771–781.

[2] M. Kosmin, J. Ledsam, B. Romera-Paredes, R. Mendes, S. Moinuddin, D. de Souza, L. Gunn, C. Kelly, C.O. Hughes, A. Karthikesalingam, C. Nutting, and R.A. Sharma. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. *Radiotherapy and Oncology*, 135:130–140, 2019.

[3] George TY Chen, Gregory C Sharp, and Shinichiro Mori. A review of image-guided radiotherapy. *Radiological physics and technology*, 2(1):1–12, 2009.

[4] Tony Lomax. State-of-the-art proton therapy: The physicist's perspective [slides]. `https://www.ptcog.ch/archive/conference_p&t&v/PTCOG47/presentations/4_Meeting_Thursday/T%20Lomax.pdf`, 2008. Last accessed 19 Mai 2022.

[5] Jean Léger, Eliott Brion, Paul Desbordes, Christophe De Vleeschouwer, John A Lee, and Benoit Macq. Cross-domain data augmentation for deep-learning-based male pelvic organ segmentation in cone beam ct. *Applied Sciences*, 10(3):1154, 2020.

[6] Seiichi Uchida. Image processing and recognition for biological images. *Development, growth differentiation*, 55, 04 2013.

[7] Anders Eklund, Paul Dufort, Mattias Villani, and Stephen LaConte. Broccoli: Software for fast fmri analysis on many-core cpus and gpus. *Frontiers in neuroinformatics*, 8:24, 2014.

[8] H. Knutsson and M. Andersson. Morphons: segmentation using elastic canvas and paint on priors. In *IEEE International Conference on Image Processing 2005*, volume 2, pages II–1226, 2005.

[9] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988.

[10] Chung-Ming Chen, Henry Horng-Shing Lu, and Yu-Chen Lin. An early vision-based snake model for ultrasound image segmentation. *Ultrasound in medicine & biology*, 26(2):273–285, 2000.

[11] Md Zahangir Alom, Tarek M Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S Awwal, and Vijayan K Asari. The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*, 2018.

[12] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9(4):611–629, 2018.

[13] Junxi Feng, Xiaohai He, Qizhi Teng, Chao Ren, Honggang Chen, and Yang Li. Reconstruction of porous media from extremely limited information using conditional generative adversarial networks. *Phys. Rev. E*, 100:033308, Sep 2019.

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[15] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, and Jose Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70:41–65, 2018.

[16] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

[17] Nicolas Audebert, Bertrand Saux, and Sébastien Lefèvre. Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 140, 11 2017.

[18] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[19] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]).

[20] Burr Settles. Active learning literature survey. 2010.

[21] J Ferlay, M Ervik, F Lam, M Comlombet, L Mery, and M Piñeros, 2020. Global Cancer Observatory: Cancer Today. Lyon: International Agency for Research on Cancer; (https://gco.iarc.fr/today, accessed February 2021).

[22] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021.

[23] G Delaney, S Jacob, C Featherstone, and M Barton. The role of radiotherapy in cancer treatment: estimating optimal utilization from a review of evidence-based clinical guidelines. 2005. Cancer, 104(6), 1129–1137.

[24] Gregory Sharp, Karl D. Fritscher, Vladimir Pekar, Marta Peroni, Nadya Shusharina, Harini Veeraraghavan, and Jinzhong Yang. Vision 20/20: Perspectives on automated image segmentation for radiotherapy. *Medical Physics*, 41(5):050902.

[25] O. Dahlqvist Leinhard, A. Johansson, J. Rydell, O. Smedby, F. Nystrom, P. Lundberg, and M. Borga. Quantitative abdominal fat estimation using mri. In *2008 19th International Conference on Pattern Recognition*, pages 1–4, 2008.

[26] Catarina Veiga, Guillaume Janssens, Ching-Ling Teng, Thomas Baudier, Lucian Hotoiu, Jamie R. McClelland, Gary Royle, Liyong Lin, Lingshu Yin, James Metz, Timothy D. Solberg, Zelig Tochner, Charles B. Simone, James McDonough, and Boon-Keng Kevin Teo. First clinical investigation of cone beam computed tomography and deformable registration for adaptive proton therapy for lung cancer. *International Journal of Radiation Oncology*Biology*Physics*, 95(1):549–559, 2016. Particle Therapy Special Edition.

[27] Christine Boydev, David Pasquier, Foued Derraz, Laurent Peyrodie, Abdelmalik Taleb-Ahmed, and Jean-Philippe Thiran. Automatic prostate segmentation in cone-beam computed tomography images using rigid registration. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3993–3997, 2013.

[28] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020.

[29] Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A. Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M. Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radiology: Artificial Intelligence*, 2(3):e190043, 2020. PMID: 32510054.

[30] M Bach Cuadra, Valérie Duay, and J-Ph Thiran. Atlas-based segmentation. In *Handbook of biomedical imaging*, pages 221–244. Springer, 2015.

[31] Jan-Jakob Sonke, Marianne Aznar, and Coen Rasch. Adaptive radiotherapy for anatomical changes. In *Seminars in radiation oncology*, volume 29, pages 245–257. Elsevier, 2019.

[32] Christos Davatzikos, Dinggang Shen, Ashraf Mohamed, and Stelios K Kyriacou. A framework for predictive modeling of anatomical deformations. *IEEE transactions on medical imaging*, 20(8):836–843, 2001.

[33] Claudio Fiorino, Franca Foppiano, Paola Franzone, Sara Broggi, Pietro Castellone, Michela Marcenaro, Riccardo Calandrino, and Giuseppe Sanguineti. Rectal and bladder motion during conformal radiotherapy after radical prostatectomy. *Radiotherapy and Oncology*, 74(2):187–195, 2005.

[34] Casey Bojechko, Patricia Hua, Whitney Sumner, Kripa Guram, Todd Atwood, and Andrew Sharabi. Adaptive replanning using cone beam ct for deformation of original ct simulation. *Journal of Medical Radiation Sciences*, 2021.

[35] G Bernard. Incremental organ segmentation with machine learning techniques : application to radiotherapy. `http://hdl.handle.net/2078.1/153438`. Prom. : Lee, John.

[36] Jasjit S Suri, David L Wilson, and Swamy Laxminarayan. *Handbook of Biomedical Image Analysis: Volume II: Segmentation Models Part B*. Springer, 2005.

[37] J.B.Antoine Maintz and Max A. Viergever. A survey of medical image registration. *Medical Image Analysis*, 2(1):1–36, 1998.

[38] Leonardo Romero and Félix Calderón. A tutorial on parametric image registration. *Scene Reconstruction, Pose Estimation and Tracking*, pages 167–184, 2007.

[39] Jerome Plumat, Mats Andersson, Guillaume Janssens, Jonathan de, Xivry, Hans Knutsson, and Benoit Macq. Image registration using morphon algorithm: an itk implementation. 04 2009.

[40] Andreas Wrangsjö, Johanna Pettersson, and Hans Knutsson. Non-rigid registration using morphons. In *Scandinavian Conference on Image Analysis*, pages 501–510. Springer, 2005.

[41] Guillaume Janssens, Laurent Jacques, Jonathan Orban de Xivry, Xavier Geets, and Benoit Macq. Diffeomorphic registration of images with variable contrast enhancement. *International journal of biomedical imaging*, 2011, 2011.

[42] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 924–931. Springer, 2006.

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

[44] Juan Eugenio Iglesias and Mert R Sabuncu. Multi-atlas segmentation of biomedical images: a survey. *Medical image analysis*, 24(1):205–219, 2015.

[45] Torsten Rohlfing, Robert Brandt, Randolf Menzel, and Calvin R Maurer Jr. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. *NeuroImage*, 21(4):1428–1442, 2004.

[46] Arno Klein, Brett Mensh, Satrajit Ghosh, Jason Tourville, and Joy Hirsch. Mindboggle: automated brain labeling with multiple atlases. *BMC medical imaging*, 5(1):1–14, 2005.

[47] Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–921, 2004.

[48] Laurent D Cohen. On active contour models and balloons. *CVGIP: Image understanding*, 53(2):211–218, 1991.

[49] Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1(2):119–130, 1988.

[50] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[51] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[52] Tomaž Vrtovec, Domen Močnik, Primož Strojan, Franjo Pernuš, and Bulat Ibragimov. Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods. *Medical Physics*, 47(9):e929–e950, 2020.

[53] Forest Agostinelli, Matthew Hoffman, Peter Sadowski, and Pierre Baldi. Learning activation functions to improve deep neural networks. *arXiv preprint arXiv:1412.6830*, 2014.

[54] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

[55] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[56] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

[57] Andreas Zell. *Simulation neuronaler netze*, volume 1. Addison-Wesley Bonn, 1994.

[58] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[59] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[60] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[61] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, 2016.

[62] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.

[63] Mostert M., Bredenoord A., Biesaart M., and al. Big data in medical research and eu data protection law: challenges to the consent or anonymise approach. *Eur J Hum Genet*, 24:956–960, 2016.

[64] Hao Jin, Yan Luo, Peilong Li, and Jomol Mathew. A review of secure and privacy-preserving medical data sharing. *IEEE Access*, 7:61656–61669, 2019.

[65] Eric B Baum and Kenneth Lang. Query learning can work poorly when a human oracle is used. In *International joint conference on neural networks*, volume 8, page 8. Beijing China, 1992.

[66] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

[67] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer, 1994.

[68] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71:102062, 2021.

[69] Andreas Vlachos. A stopping criterion for active learning. *Computer Speech & Language*, 22(3):295–312, 2008.

[70] Punit Kumar and Atul Gupta. Active learning query strategies for classification, regression, and clustering: a survey. *Journal of Computer Science and Technology*, 35(4):913–945, 2020.

[71] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. *Advances in neural information processing systems*, 23, 2010.

[72] Steven CH Hoi, Rong Jin, Jianke Zhu, and Michael R Lyu. Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd international conference on Machine learning*, pages 417–424, 2006.

[73] Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, 2:441–448, 2001.

[74] Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. *Advances in neural information processing systems*, 20, 2007.

[75] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pages 309–318. Springer, 2001.

[76] Alex Holub, Pietro Perona, and Michael C Burl. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008.

[77] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

[78] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.

[79] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992.

[80] Ion Muslea, Steven Minton, and Craig A Knoblock. Selective sampling with redundant views. In *AAAI/IAAI*, pages 621–626, 2000.

[81] William H. Beluch, Tim Genewein, Andreas Nurnberger, and Jan M. Kohler. The power of ensembles for active learning in image classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9368–9377, 2018.

[82] Jeroen Bertels, Tom Eelbode, Maxim Berman, Dirk Vandermeulen, Frederik Maes, Raf Bisschops, and Matthew B Blaschko. Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 92–100. Springer, 2019.

[83] Eliott Brion, Jean Léger, A.M. Barragán-Montero, Nicolas Meert, John A. Lee, and Benoit Macq. Domain adversarial networks and intensity-based data augmentation for male pelvic organ segmentation in cone beam ct. *Computers in Biology and Medicine*, 131:104269, 2021.

[84] Lu Liu, Jelmer Maarten Wolterink, Christoph Brune, and Raymond NJ Veldhuis. Anatomy-aided deep learning for medical image segmentation: a review. *Physics in Medicine & Biology*, 2021.

[85] Murali Shanker, Michael Y Hu, and Ming S Hung. Effect of data standardization on neural network training. *Omega*, 24(4):385–397, 1996.

[86] Ju Xu, Mengzhang Li, and Zhanxing Zhu. Automatic data augmentation for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 378–387. Springer, 2020.

[87] Eliott Brion, Jean Léger, Umair Javaid, John Lee, Christophe De Vleeschouwer, and Benoit Macq. Using planning cts to enhance cnn-based bladder segmentation on cone beam ct. In *Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 10951, pages 410–417. SPIE, 2019.

[88] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[89] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[90] Luke Metz, Niru Maheswaranathan, Jeremy Nixon, Daniel Freeman, and Jascha Sohl-Dickstein. Understanding and correcting pathologies in the training of learned optimizers. In *International Conference on Machine Learning*, pages 4556–4565. PMLR, 2019.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain
Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/epl**