Data Augmentation to Improve Cross-Domain Generalization in Deep Learning MRI Segmentation

Rasmus Helander

Master's thesis
2021:E26

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematics

CENTRUM SCIENTIARUM MATHEMATICARUM

Abstract

Semantic segmentation of medical images is an important task with many applications. However, manually delineating 3D images is time-consuming and the demand for automation is high. For many image segmentation tasks, deep learning has provided state-of-the-art results. However, the varying nature of magnetic resonance (MR) images due to factors such as different machine vendors and clinical protocols has been found to lead to domain transfer issues for deep learning segmentation models.

This thesis aims to investigate three different data augmentation methods to remedy this domain generalization problem. We first consider a standard data augmentation approach and study the effects of applying stacked image processing functions to the training data. We then study the effects of including weakly labeled training samples from the unseen data domain in the training set. Lastly we study the effects of training a Cycle GAN to transfer labeled training samples to the unseen domain, and including this synthetic data in the training. The experiments are carried out on MRI data from three separate domains, where labeled training samples only exist for one of them. The studied methods are found to increase average DICE scores on the rectum and urinary bladder by 9-14% and 25-40% on the two unseen domains, and 95th percentile Hausdorff distances are decreased by 60-71% and 33-54%. Additionally, combining the use of unlabeled data from unseen domains with extensive image processing is found to further boost segmentation performance.

Acknowledgements

I would like to thank my supervisors; Alexandros Sopasakis from LTH and Jonas Söderberg from RaySearch Laboratories. Throughout the course of this thesis, they have provided valuable insights and comments to drive the work forward. I would also like to thank everyone else at RaySearch who has been a part of this project in some way, as well as my fellow master's thesis students at the company - particularly Jonas Berg, not only for relevant discussions and code-sharing but also for valuable face-to-face socializing during the work-at-home era of the Covid-19 pandemic. At RaySearch, I would finally like to direct special thanks to David Lidberg who has been a regular participant in weekly meetings, and who's knowledge within medical image segmentation has been of great help.

Contents

1 Introduction

1.1 Motivation

This thesis was carried out at RaySearch Laboratories, a medical software company specialised in the generation of radiotherapy treatment plans for cancer patients. In the creation of such treatment plans, the segmentation of regions of interest (RoI:s) - such as organs or tumours - in a patient image is a crucial step. Furhermore, the applications of medical image segmentation go beyond treatment planning and include areas such as diagnostics and population level research [1]. However, manually detecting and delineating RoI:s in 3D images is tedious and time-consuming work, leading to a demand for automation. This could also remove the risk of error due to factors such as fatigue and reduce variations between different practitioners.

Deep learning is a commonly used technique for medical image segmentation and is available in several treatment planning systems for radiotherapy such as RayStation, where a common use case is segmentation of organs at risk on Computed Tomography (CT) scans. This thesis instead focuses on MR images, which have several benefits over CT - for example, they are significantly better at differentiating soft tissue and do not require the patient to be exposed to ionizing radiation [2]. These benefits have led to the suggestion of an MRI-only treatment planning workflow (the current workflow requires CT images to generate the treatment plans), which has received high interest [3, 4] and motivates the development of accurate MR segmentation models. MRI-only planning aside, the current CT-based treatment-planning workflow relies on image registration between CT and MR images, for which MR segmentation is also crucial [3].

1.2 Challenges

Unfortunately, training MR segmentation models is not an easy task due to several factors. Firstly, as with many other medical image analysis problems, acquiring sufficient amounts of high-quality annotated training data is not an easy (or cheap) task. Secondly, there is a large variation apparent in MR images due to factors such as differences in the hardware of the image acquiring machines, scanning settings and clinical protocols [1]. The first difficulty is present for the case of CT scans too, but the second is more unique to MR. The practical implication of this second problem is that a segmentation model that is trained on data from one clinic (interpreted as one data domain) often experiences significant performance loss when used on data from another clinic. The simple solution to this would be to include data from many different clinics in the training of the model, but the scarcity of high-quality annotations makes this approach infeasible, in practice if not in theory. Another option, which is also the approach chosen for this thesis, is to use data augmentation in order to reduce the impact of domain shift.

1.3 Aim

The aim of this thesis is to analyze several methods of performing data augmentation for training deep learning segmentation models for MR images. Specifically, we aim to answer the following questions:

- Is it possible to reduce the negative impact of domain transfer on segmentation model performance by applying data augmentation at training-time?
- Of the methods tested, which is the most appropriate one?
- Can anything be said about what domain differences in MR images are contributing most to domain shift performance loss?

2 Background

2.1 Medical image segmentation

In general, image segmentation refers to the task of automatically partitioning an image into meaningful subparts, or segments. In medical imaging, these RoI:s are often specific organs or pathologies such as cancer tumours, and the segmented images can be used for several clinical applications such as diagnostics, treatment planning and patient follow-up [5]. Figure 1 shows a few examples of images of the male pelvic region and corresponding label masks for the rectum and the urinary bladder.

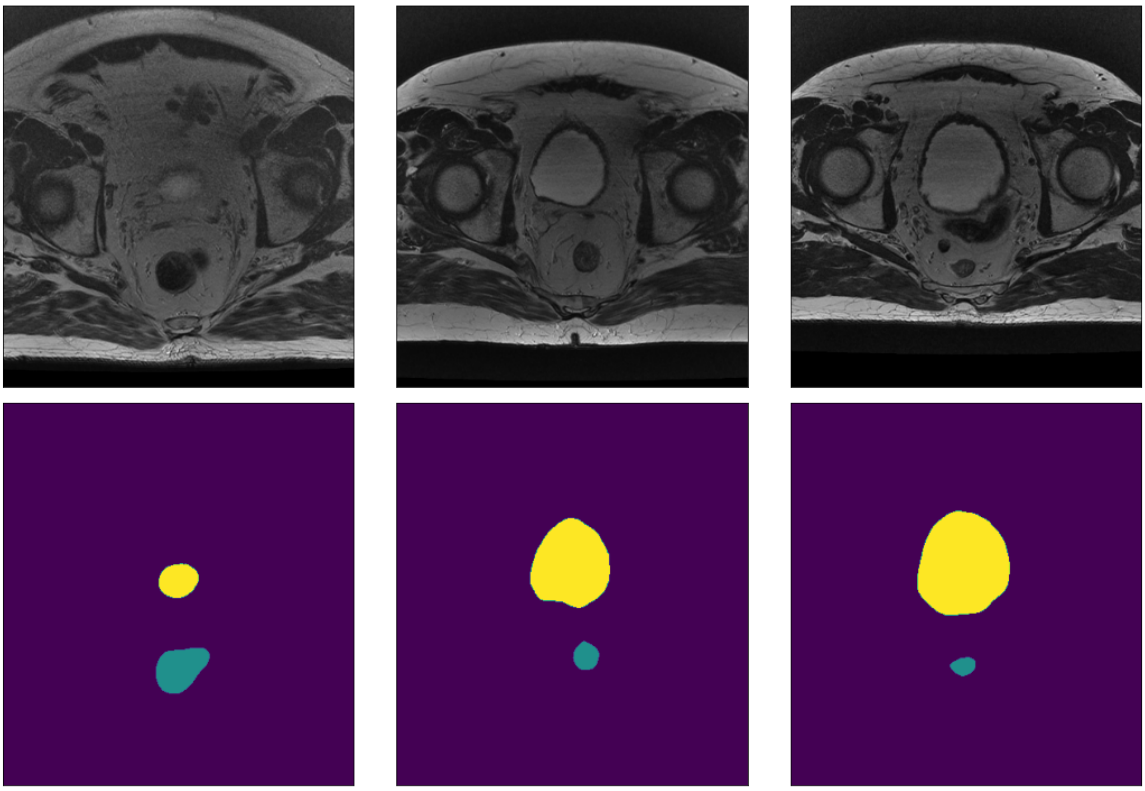


Figure 1: Slices of pelvic MR images and their corresponding label masks for the urinary bladder (yellow) and the rectum (green). Background voxels are assigned a value of 0 in the mask.

To formalize and provide some notation, the task of medical segmentation is the following: given an image (in 2 or 3 dimensions) $\mathbf{x}$ of a patient's anatomy and a set of RoI:s, finding the mask $\mathbf{y}$ that corresponds to the locations of the RoI:s in the image - in other words finding a mapping f such that

$$f(\mathbf{x}) = \mathbf{y}. \tag{1}$$

It is convenient to think of this as a set of classification tasks: for each pixel (voxel, in the case of 3D images) in $\mathbf{x}$, the goal is to determine which RoI the pixel belongs to (or rather which RoI the majority of the pixel belongs to, seeing as the physiological reality of the patient is not divided into pixels).

There currently exists a number of different automated segmentation techniques [5], but this thesis is focused on a deep learning based method discussed in Section 3 for finding f. For this method, the task is formulated as utilizing a set of n training images $X = \{\mathbf{x}_i\}_1^n$ and corresponding hand-annotated RoI masks $Y = \{\mathbf{y}_i\}_1^n$ to find a function f, parameterized by a deep neural network, that minimizes the distance between a predicted mask $\hat{\mathbf{y}} = f(\mathbf{x})$ and the true RoI mask $\mathbf{y}$. Specifically, one aims to find a function that generalizes well to images that are not contained in the training set - in other words, one seeks to find

$$f^* = \arg\min_{f \in H} \mathbb{E}_{(\mathbf{x},\mathbf{y}) \in \mathcal{X} \times \mathcal{Y}} \left[D\left(f\left(\mathbf{x}\right), \mathbf{y} \right) \right] \tag{2}$$

for some distance measure D, where H represents the hypothesis space (i.e. the parameter space of the neural network) and where the expectation is taken over the joint space of images and labels $\mathcal{X} \times \mathcal{Y}$. As with any deep learning model, one tries to find f^* by optimizing an objective (consisting of a loss function and some regularizers) on the training data X and Y, by means of gradient descent or some stochastic version thereof.

2.2 Domain transfer - concepts and notation

As mentioned, one of the prominent challenges within medical image segmentation tasks is scarcity of high-quality labeled data [6]. Deep learning models for natural images are usually trained on large data sets, and limited resources in terms of time, money and expertise makes it challenging to curate medical image data sets of similar size. This means that the problem of deep learning models overfitting to the training set is an important consideration when developing computer vision models for medical image segmentation. Furthermore, differences in the setup used when acquiring the images (such as different modalities, scanning machines and clinical protocols) as well as populational variations between e.g. different countries provides an additional challenge to medical image segmentation tasks called *domain shift* [7]. Put in words, the domain shift problem means that *the performance of a deep learning model trained on one set of images might be significantly worsened when tested on a different set of images.*

Mathematically, this can be understood as a space $\mathcal{X} \times \mathcal{Y}$ of features and labels, with different probability distributions P_S and P_T representing the *source* and the *target* domains, respectively. Suppose then that we have a set $\mathcal{D}_S = \{(\mathbf{x}_i^S, \mathbf{y}_i^S)\}_{i=1}^{n_S}$ of n_S samples drawn from $\mathcal{X} \times \mathcal{Y}$ with probability P_S and a set $\mathcal{D}_T = \{(\mathbf{x}_i^T, \mathbf{y}_i^T)\}_{i=1}^{n_T}$ of n_T samples drawn with probability P_T. A deep learning model m is trained on a training subset $\mathcal{D}_S^{train} \subset \mathcal{D}_S$ and tested on a testing subset $\mathcal{D}_S^{test} \subset \mathcal{D}_S \setminus \mathcal{D}_S^{train}$ from the source domain as well as on a subset $\mathcal{D}_T^{test} \subset \mathcal{D}_T$. Using some performance metric P, we can now define the *domain shift error* as the expected performance difference between the source and target domains when m is used

$$\epsilon_{DS} := \mathbb{E}_{\mathcal{D}_S^{test}, \mathcal{D}_T^{test}}(P(\mathbf{x}^S, \mathbf{y}^S, m) - P(\mathbf{x}^T, \mathbf{y}^T, m)). \tag{3}$$

Naturally, the notation above can be expanded to include an arbitrary number of source and/or target domains.

An intuitive solution to the domain shift problem would be to include data from both source and target domains in the training of the deep learning model. However, this is often unfeasible due to the scarcity of labeled training data discussed above. Instead, a

number of other solutions have been proposed to the problem. Guan and Liu [7] - from whom a lot of the notation in this report is borrowed - provide a detailed overview of such *domain adaptation* approaches used for segmentation as well as other medical image analysis tasks. They divide deep learning approaches for the problem into the following four categories:

- **Supervised domain adaptation.** When a limited amount of labeled examples from the target domain is available for training
- **Semi-supervised domain adaptation.** When in addition to a small amount of labeled examples, redundant *unlabeled* data from the target domain is available for training
- **Unsupervised domain adaptation.** When only unlabeled data from the target domain is available for training
- **Domain generalization** When no target data is available for training, neither images nor labels.

In addition to the above categories, one can add the following:

- **Weakly supervised domain adaptation** When no labeled target data is available, but one can obtain noisy, or weak, target labels with little or no effort

For the purpose of this thesis, unsupervised and weakly supervised domain adaptation is investigated alongside domain generalization.

2.3 Magnetic resonance imaging

The workings of an MRI scanner is a joint feat of modern physics, mathematics and engineering, and the details of the technique are far beyond the scope of this thesis. However, this section gives an overview of some of the important concepts so that the reader understands what is seen in an MR image and some sources of variation. The overview is, unless otherwise indicated, based on *MRI - From Picture to Proton* by McRobbie et. al [8].

2.3.1 Generating the MR signal

The fundamental reason that MR phenomena are observed is the *magnetic moment of atomic nuclei*, which is a result of the quantum mechanical interactions between the angular momenta of the nucleons (protons and neutrons) [9]. This moment is intrinsically connected to the spin of the nucleus [10] which means that it is only non-zero for atoms where either the number of protons or the number of neutrons (or both) is odd [9], such as for the hydrogen atom, the nucleus of which consists of a single proton.

What happens in an MR scanner, is that a sample (i.e. the body of the patient) is placed in an external magnetic field. Using the terms of classical physics, this forces the magnetic moments of the nuclei in the sample to align with the external field. Quantum mechanically, the external field provides a z axis to which the nuclear spins can either align parallelly (spin up) or antiparallelly (spin down). These two states have an energy difference proportional to the strength of the external field which makes the population of the spin up state slightly larger than that of the spin down state, leading to a *net magnetization* of the sample parallel to the external field. However, the field of the net magnetization is too weak to be measured, compared to the external field.

In order to get a measurable signal, one applies a *radiofrequency* (RF) pulse to the patient, i.e. a short pulse of electro-magnetic radiation. If the frequency of the RF pulse matches the energy gap between the spin up and down states, it stimulates the nuclei to flip from one state to the other - the duration of the pulse determines the probability to switch states. Without going into the details, one can ensure that the RF pulse changes the direction of the net magnetization of the sample, from the z axis (as defined by the external magnetic field) to the xy-plane, where the field direction will oscillate with the frequency of the RF pulse. As the net magnetization of the sample is now perpendicular to the external field, it can be measured - this is the MR signal.

2.3.2 Contrast in MR images

The above explanation provides a description of how the MR signal is generated, but not why it is an interesting quantity for medical purposes. The reason for this is that different kinds of tissue have properties that in several ways effect the MR signal. There are three such factors that are of great importance for the resulting image; proton density (PD), spin-lattice relaxation time (T1) and spin-spin relaxation time (T2).

The first, PD, is somewhat self explanatory. The measured signal will naturally depend on the density of the nuclei that are emitting the signal. This means that the measured signal will be high for tissue that contains high levels of hydrogen, essentially tissue with high water content. The two latter, T1 and T2, are not as intuitive to understand but are both connected to the decay in signal over time. As stated above, the interaction between the RF pulse and the nuclei of the sample causes the net magnetization to dis-align with the external magnetic field, generating a measurable field in the perpendicular plane. However, this *transversal* magnetic field will eventually disappear, due to two separate phenomena in the nuclei. Firstly, the external magnetic field will cause the net magnetization to realign with the direction of the field, as excited nuclei go back to the spin up state by interacting with (losing energy to) its surrounding environment. This process is called spin-lattice relaxation, and T1 is the time constant related to this decay in signal. Secondly, the reason that there is a field in the xy-plane is not only that the net magnetization vector of the sample is 'tilted down', but also that the individual magnetic moments of the nuclei oscillate in phase with eachother. Over time, interaction between a nucleus and its microscopic environment (e.g. other nuclei and electrons) causes this precession to dephase, eventually netting out to zero. This is called spin-spin relaxation, and T2 is the related time constant. Note that unlike spin-lattice relaxation, spin-spin relaxation has nothing to do with the realignment of the net magnetization to the external

field, and is thus in essence not affected by the external field strength.

Differences in PD, T1 and T2 is what makes the signal vary between different tissue, creating the contrast of the MR image. By using a series of different RF pulses (pulse sequences), and applying gradients in the external magnetic field (that is, varying the field strength along the spatial axes), it is possible to generate images where the contrast (difference in signal between various tissue) stems from T1 properties, T2 properties or PD. To no surprise, these are called T1-weighted, T2-weighted and PD-weighted, respectively. The reader should note that these types of images are not mutually exclusive categories - the measured contrast will always be a result of differences in both transversal and longitudinal relaxation, as well as proton density. However, with different pulse sequences one can manipulate the atomic interactions with the magnetic field so that T1, T2 or PD properties are either highlighted or suppressed. Two properties that are important for determining whether the image will be T1- or T2-weighted are the *time to repetition* (TR) of the pulse sequence, and the *time to echo* (TE), which can be interpreted as the time between initialization of a pulse sequence and the measurement of the signal. In general, T1-weighted images have short TR and TE while T2-weighted images have longer TR and TE.

2.3.3 Sources of variation

To summarize, there are a number of factors that effect MR image differences, besides inter-patient variation. Firstly, the strength of the magnetic field (and the TE) will have a great impact on the general signal strength of the image. Secondly, variations in parameters such as the choice of pulse sequence, TR and TE will will result in differences in tissue contrast. For example, in T1-weighted images the signal of fat will be stronger than the signal of fluids, which means that fat will appear brighter in the image. On the other hand, the signal from fluid is in general higher than the one from fat in T2-weighted images.

In addition to variations in the signal strength, different scanner settings will lead to different signal to noise-ratios, which is also affected by choices in the image-reconstruction process, i.e. the process of turning the measured signal into an image. Furthermore, MR images can contain a number of artefacts which are also affected by the choice of imaging parameters.

3 Data

In this thesis, labeled and unlabeled 3D MRI scans of the male pelvic region from three different data sets are used to investigate the impact of data augmentation on the domain shift problem. The RoI:s studied are the urinary bladder and the rectum. The three data sets differ in several ways including scanner, choice of pulse sequence and weighting, and are described below. A summary of the data sets are given in Table 1 and a more thourough description can be found from the specified reference for each dataset.

Table 1: Information regarding the MR scanner settings used to create the data sets. For $\mathcal{D}_{T2}$, most information about the dataset is unavailable.

Domain	$\mathcal{D}_S$	$\mathcal{D}_{T1_1}$	$\mathcal{D}_{T1_2}$	$\mathcal{D}_{T1_3}$	$\mathcal{D}_{T2}$
Pulse sequence	2D FRFSE	FRFSE	TSE	FRFSE	SE
Scanner	GE Discovery 750 W 3.0 T	Unknown	Unknown	Unknown	Siemens Skyra$_{fit}$
TE (ms)	96	97	91-102	65	Unknown
TR (ms)	15 000	6000-6600	12000-16000	9000	Unknown
Band width per pixel (Hz)	390	390	200	390	Unknown
Slice thickness (mm)	2.5	2.5	2.5	2.5	4
Voxel size (mm^3)	0.44×0.44	0.875×0.875	0.875×0.875-$2.5 \times 1.1 \times 1.1$	0.875×0.875	0.875×0.875
Reference	[5]	[11]	[11]	[11]	N/A

The data were exported in mhd-format from RayStation. For the labeled images, a box centered around the bladder and the rectum was exported. For images with no corresponding labels, a mask separating the patient's body from the surroundings was generated by thresholding and a box centered around the center of the generated mask was exported. Admittedly, there is some risk to this procedure as it does not ensure that the two kinds of images are geometrically identical. However, as the bladder and rectum are both situated at the center of the scans this was deemed acceptable. Furthermore, ocular inspection of the two types of procedure did not show significant differences.

The volumes used were all of size $388 \times 421 \times 114$ voxels, and the spatial resolution was $2.5 \times 0.7 \times 0.7$ mm. As is apparent from Table 1, not all images were acquired with this resolution, meaning that the more coarse-grained images were interpolated to fit the resolution of the remaining scans.

3.1 Source Domain - $\mathcal{D}_S$

The source domain is represented by a set of T2-weighted images from the MR-PROTECT trial [4], an MRI-only treatment planning study in Lund, Sweden. This data was made available for this thesis through the national ASSIST project, a joint project between Swedish healthcare institutions, academia and industry with the aim of collecting and processing clinical radiotherapy data for machine learning. From this image set, a total of 38 images and corresponding manually catered labels are used with the following split; 24 train, 6 validation and 8 test.

3.2 Target Domain 1 - $\mathcal{D}_{T1}$

The first target domain consists of images and labels from the Gold Atlas project [11], publicly available at *https://zenodo.org/record/583096.YIbi1y0Rq3U*. The dataset originally consists of 19 scans of the male pelvic, of which 18 were used (one was removed due to the patient having an artificial femur). These 18 images were divided into validation and test sets with 9 images each. As indicated by Table 1, the images of $\mathcal{D}_{T1}$ are acquired with 3 different scanners and different settings, meaning that there exists distinct subdomains within $\mathcal{D}_{T1}$. The RoI delineations are created by 5 human practitioners agreeing on a consensus.

3.3 Target Domain 2 - $\mathcal{D}_{T2}$

For the second target domain, a subset of a large unlabeled dataset of anonymized MR scans from Iridium Kankernetwerk, Antwerp, Belgium was used. 19 images were selected from the data set and were manually labeled by myself under the supervision of a clinical expert. These images were used for validation (9 samples) and testing (10 samples). In addition to these, 52 unlabeled images were used for the experiments which rely on unlabeled target data.

3.4 Domain differences and data set statistics

To give a sense of how the three data domains differ from one another, Figure 2 shows sample slices from each of the domains. From these images it is apparent that the MR scans vary in several ways, such as mean intensity level, tissue contrast, RoI volume, patient size, RoI homogeneity as well as the vertical placement of the RoI:s (for example, the bladder in the bottom middle scan has not yet started at the specific slice). It is also striking that there is a variation in these features *within each domain* as well, for example the relative intensity in the bladder is significantly higher in the top scan from $\mathcal{D}_{T2}$ than in the bottom scan. Figure 3 shows a zoomed-in view of the slices in Figure 2. Here, we see some more fine-grained differences between the domains, such as noise level and smoothness of the edges in the scans.

To quantify the domain differences, Figures 4 through 6 show scatter plots of statistics for the different domains. Figure 4 shows RoI volumes of the different domains, and compares the target domains to the source. Notice that $\mathcal{D}_{T2}$ has a much higher variance in both rectum and urinary bladder volume. Figure 5 shows the mean intensities of the RoI:s of the different domains, after standardising each image to a mean of 0 and standard deviation of 1, i.e. the standardised RoI mean intensities relative to the image means. Figure 6 shows the two first principal components of the intensity histograms of the normalized images, to further highlight the variation in intensity distributions between domains.

All of the statistics in Figures 4-6 were calculated on volumes that were cropped and centered around the urinarly bladder and rectum, as in Figure 2, as these were the scans

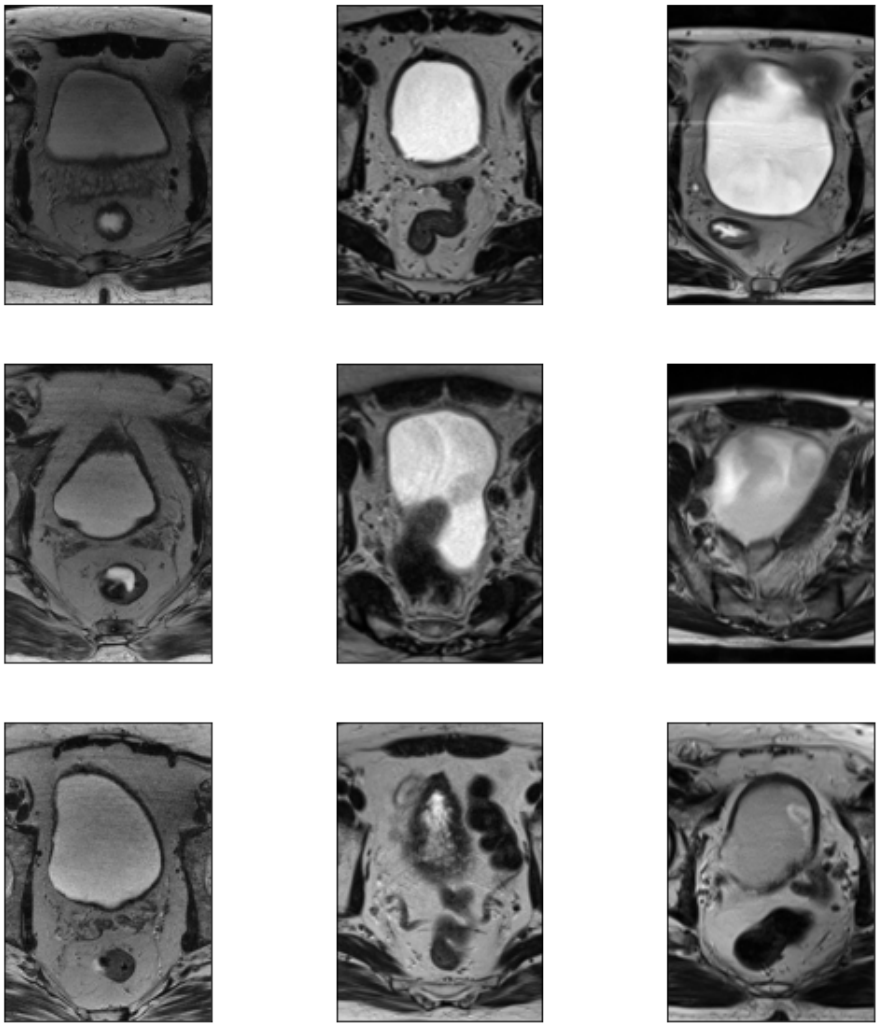


Figure 2: Slices from 9 different patients from $\mathcal{D}_S$ (left column), $\mathcal{D}_{T1}$ (middle column) and $\mathcal{D}_{T2}$ (left column). All slices show the scan in the xy-plane, with $z = 45$, and are cropped to show the bladder and the rectum of the patient. The variations between (and within) domains that are visible to the naked eye include voxel intensity distributions, noise level and RoI size.

that were used to evaluate model performance.

A source of inter-domain variation in the MR images which is not discussed here is the amount of artefacts present in the images. There are a number of artefacts that can be showcased in MR images, and some of these are more common for certain pulse sequences than for others [8]. However, correctly observing these artefacts is a detailed task and

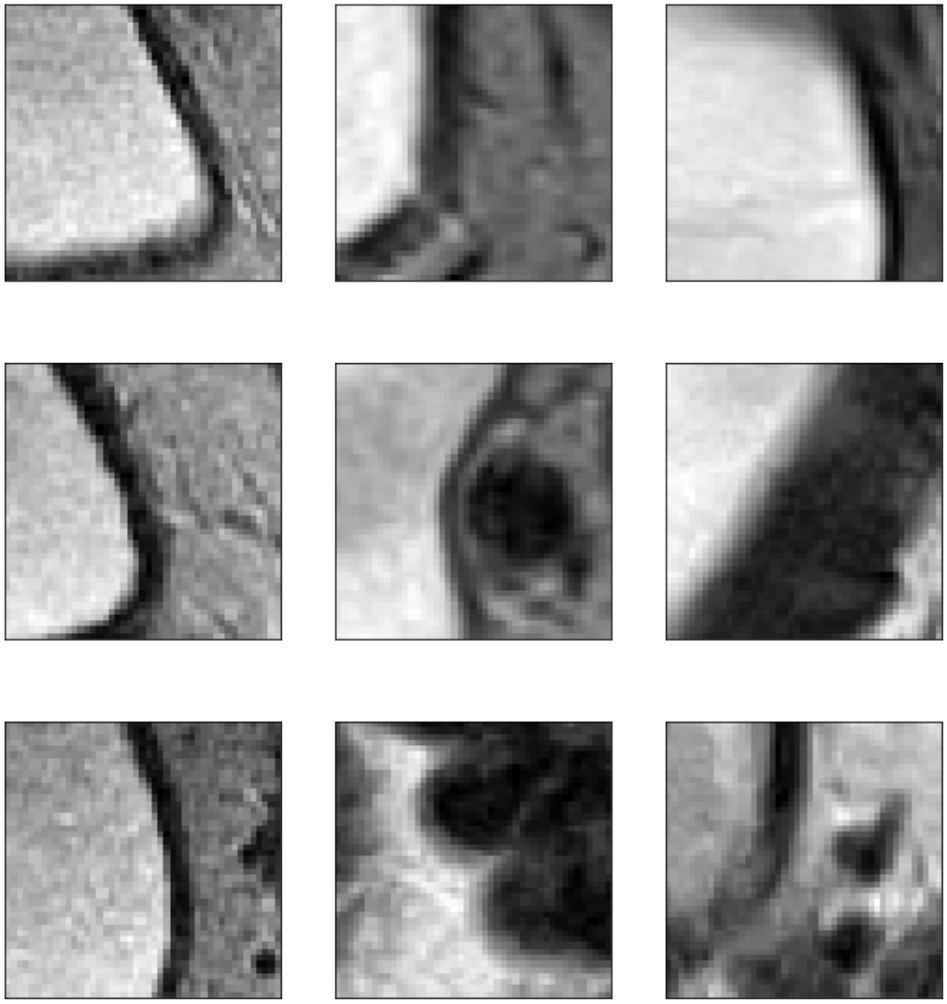


Figure 3: Slices from 9 different patients in $\mathcal{D}_S$ (left column), $\mathcal{D}_{T1}$ (middle column) and $\mathcal{D}_{T2}$ (left column). These slices are the same as in Figure 2 but zoomed in. The bladder and bladder wall (to the left in the images) are clearly visible in all scans except the bottom middle one. The $\mathcal{D}_S$ images seem to be sharper as well as contain more high-frequency noise, while the $\mathcal{D}_{T1}$ and $\mathcal{D}_{T2}$ seem more blurred.

was therefore decided to be left out of the scope for this thesis.

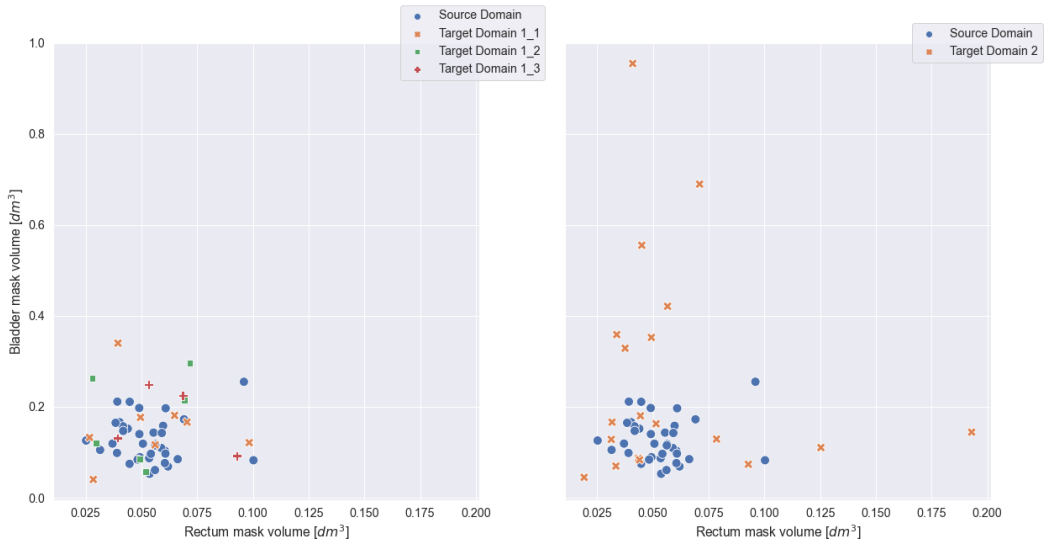


Figure 4: Rectum and bladder volume for the different domains. The distributions of $\mathcal{D}_S$ and $\mathcal{D}_{T1}$ seem to be quite similar, while $\mathcal{D}_{T2}$ includes a considerable number of samples with larger urinary bladders, and a few samples with larger rectums. These differences might be related to differences in clinical protocol.

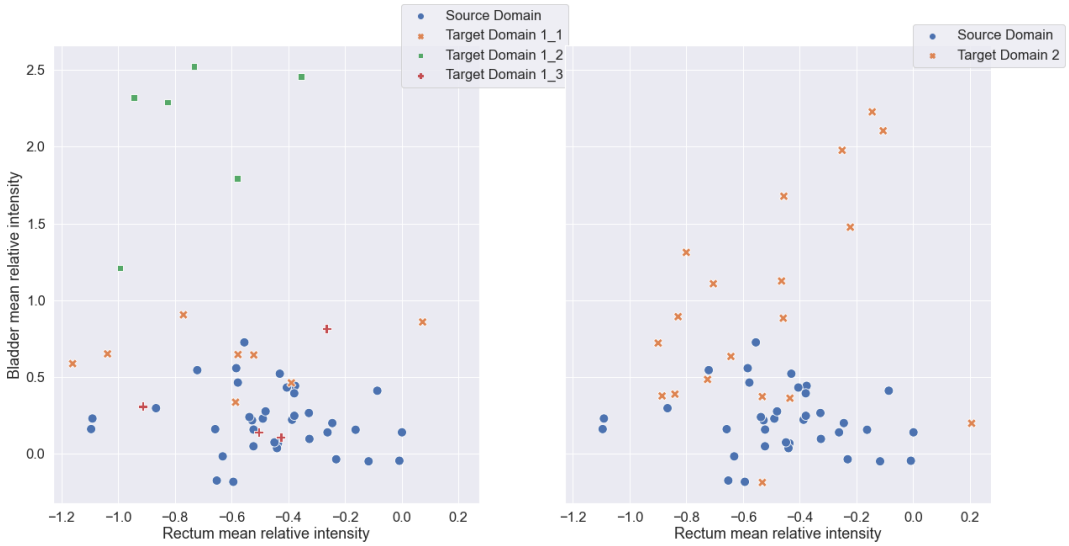


Figure 5: Rectum and bladder mean intensities for the different domains. It is clear that $\mathcal{D}_{T1_2}$ and $\mathcal{D}_{T2}$ are different from $\mathcal{D}_S$, with most samples having significantly higher relative bladder intensities. Some differences in the bladder intensities are also apparent between $\mathcal{D}_{T1_1}$ and $\mathcal{D}_S$. There does not seem to be as a significant difference between the domains in terms of the mean intensities of the rectum

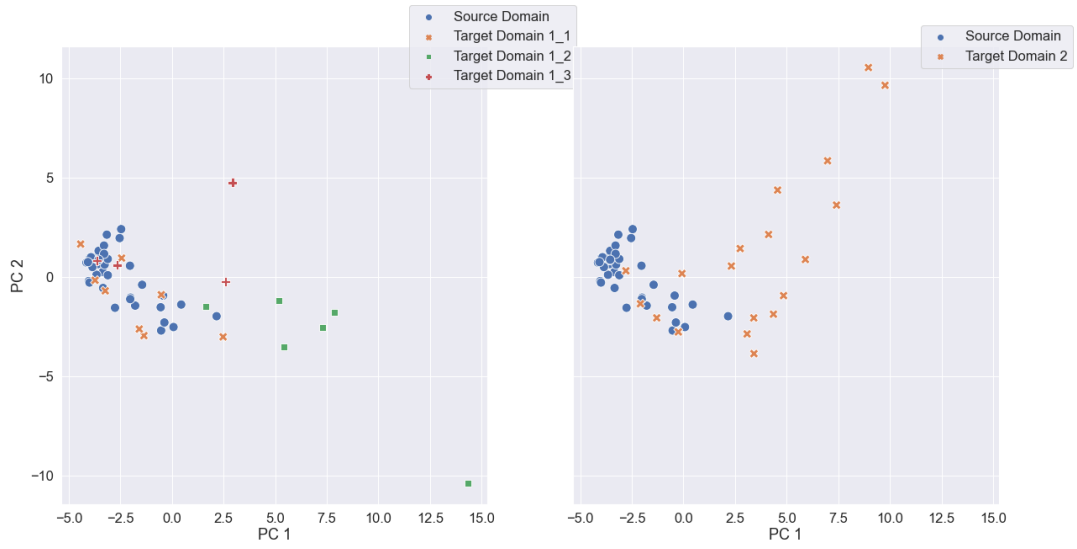


Figure 6: First two principal components of the intensity histograms of the normalized images from the different domains. It is clear that the histograms of $\mathcal{D}_{T1_2}$ and $\mathcal{D}_{T2}$ are differently distributed than those of $\mathcal{D}_S$, while $\mathcal{D}_{T1_1}$ and $\mathcal{D}_S$ seem rather similar in this aspect. The small number of samples in $\mathcal{D}_{T1_3}$ make it hard to say anything in general about its histogram distribution. All histograms were computed using 50 bins.

3.5 A note on image padding

All data domains do not cover the same physical space, i.e. the raw MR scans do not take up the same volume. For the training of deep learning models, it is necessary that all inputs are of the same shape (in voxel space), and it was decided to let one voxel represent the same physical dimensions across all domains. This means that some of the MR scans needed to be padded as part of the preprocessing of images. The choice of padding strategy might be of importance for the performance of the deep learning model, as it has a major impact on the distribution of intensities after normalization, and the need for padding varies between domains - see Figure 7.

The padding strategy used in this thesis was padding with a constant value, and no extensive experimentation was done on different padding values. However, it was found empirically that min-padding (i.e. padding with the smallest measured intensity) performed better on the baseline than padding with -1000, which is the RayStation default, indicating the impact that padding strategy has on model performance. The statistics of Figures 5 and 6 are calculated on the min-padded images, i.e. the exact same images that are fed to the neural network during training and inference.

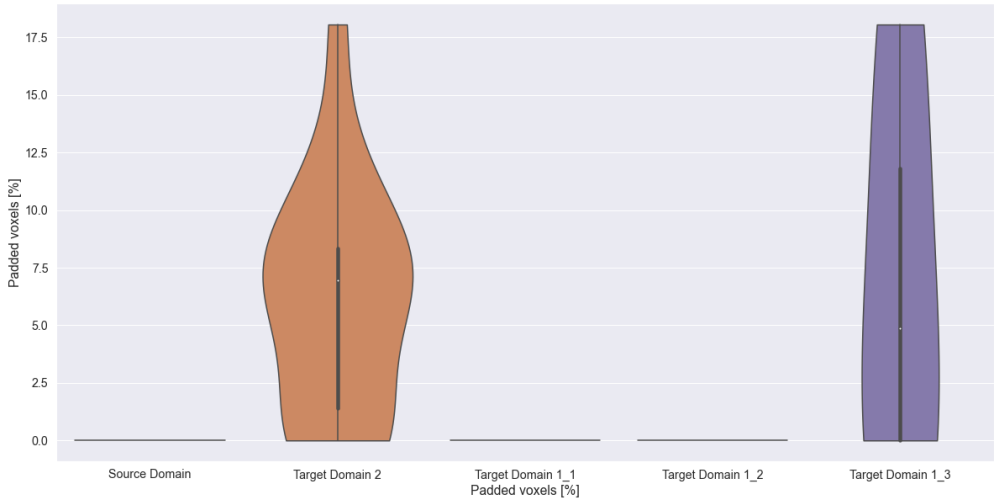


Figure 7: The distribution of padding voxel percentage over the different domains. Scans from $\mathcal{D}_{T1_2}$ and $\mathcal{D}_{T2}$ are aquired with a smaller scan region that the other domains, and thus need padding to have the same size before feeding them to the neural network. The padding of images leads to differences in intensity distributions after applying normalization or standarization.

4 Methodology

4.1 Segmentation Network

The base segmentation network used for all experiments was an in house implementation of the well known U-Net architecture, specifically developed for segmentation of medical images [12] - see Figure 8 for the general network architecture. The main difference between the original network and the used one was that it was adapted to work with 3D data instead of 2D, meaning all convolutions were in 3D.

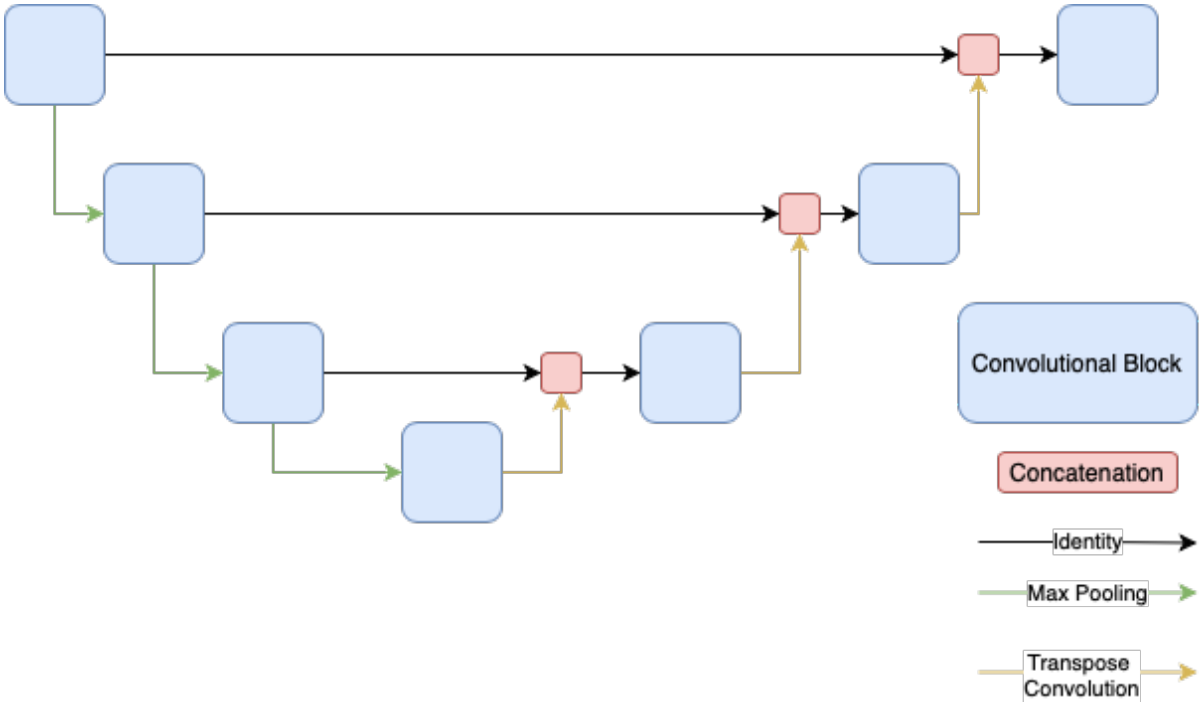


Figure 8: Neural network architecture for the U-Net

For the baseline experiments, the only preprocessing done was that the volumes were cropped to a size of $252 \times 182 \times 72$ voxels (to fit the GPU memory while still containing all of the ROI:s) and subsequently standardised to a mean of 0 and standard deviation of 1.

4.1.1 Loss function and optimization

The segmentation network outputs an array of shape $w \times d \times h \times (n_c + 1)$ where $w \times d \times h$ is the size of the input data and n_c is the number of classes (i.e. the number of RoI:s). These values are interpreted as class probabilities by applying the softmax function

$$\hat{p}_{vc} = \frac{e^{\hat{y}_{vc}}}{\sum_{c=0}^{n_c} e^{\hat{y}_{vc}}} \tag{4}$$

where $\hat{p}_{vc}$ is the probability that voxel v belongs to class c (where 0 is no class/external) and $\hat{y}_{vc}$ the corresponding value in the output array.

The loss of the segmentation network is calculated using the negative crossentropy between the ground truth labels and the predicted probabilites:

$$\mathcal{L} = -\frac{1}{|V|} \sum_{v \in V} \sum_{c=0}^{n_c} I_c(v) \log(\hat{p}_{vc}) \tag{5}$$

i.e. the mean crossentropy taken over all voxels in the output volume V. $I_c(v)$ is here representing the indicator function for the output class of voxel v.

The segmentation network is trained using the Adam Optimizer [13] with a learning rate of 10^{-3}. The number of epochs and iterations varied somewhat between experiments and are therefore specified in their respective section. Due to GPU memory constraints, the network as trained using a batch size of 1, meaning that the loss function of Equation 5 was calculated, differentiated and used to update network weights for one image at the time.

4.2 Augmentation methods

4.2.1 Traditional Data Augmentation

It is well known that using data augmentation is a powerful tool to reduce overfitting in deep learning models, especially for training setups with small data sizes [14]. A common explanation for this is that by performing data augmentation, i.e. altering the training data, the neural network will get experience from a larger portion of the possible data points than from only training on non-augmented examples [14]. This naturally leads to the question of whether it is possible to synthesize training examples from the target domains (or at least closer to them) by applying augmentation to the source domain examples. Zhang et. al. argue that this is precisely the case [15], and suggest an augmentation pipeline for the task of segmenting MR images, called BigAug.

The data augmentation experiments closely follow the pipeline of Zhang et. al. which includes the following 9 steps:

1. Blurring of the input image by convolving the image with a Gaussian kernel k with mean 0 and standard deviation σ to form $I_{blurred} = k(\sigma) * I_{input}$, where $\sigma \sim \mathcal{U}(0.25, 1)$

2. Random sharpening of the image by applying a Gaussian kernel, and then adding the difference between the blurred image (scaled by a factor α) to the input image in order to amplify high-frequency information, i.e

$$I_{sharpened} = I_{input} + \alpha(I_{input} - k(\sigma) * I_{input}) \tag{6}$$

   where $\alpha \sim \mathcal{U}(10,\ 30)$ and $\sigma \sim \mathcal{U}(0.25,\ 1)$

3. Adding Gaussian noise with mean 0 and standard deviation σ to the input image, where $\sigma \sim \mathcal{U}(0.1,\ 1.0)$

4. Shifting the intensities of the input image by some value m, where $m \sim \mathcal{U}(-0.1, 0.1)$

5. Applying gamma correction to the input image, i.e. for every voxel intensity apply the transform

$$v_{out} = (v_{in})^{\gamma} \tag{7}$$

   where $\gamma \sim \mathcal{U}(0.5,\ 4.5)$. The intensities of the input image are first normalized to lie between 0 and 1

6. Rescaling the intensities of the input image by a factor k and subsequently shifting by a factor m, i.e. applying the transformation

$$v_{out} = kv_{in} + m \tag{8}$$

   for every voxel v_{in} in the input image, where $k \sim \mathcal{U}(0.9,\ 1.1)$ and $m \sim \mathcal{U}(-0.1,\ 0.1)$

7. Rotate the image by an angle d, where $d \sim \mathcal{U}(-20°,\ 20°)$. Rotation is performed around all axes with independently sampled angles

8. Deform the image elastically by applying a smooth deformation field, generated by deforming a coarse grained (voxel size $8 \times 8 \times 8$ cm) grid with random values $\sigma \sim \mathcal{N}(0,\ 0.5)$[cm] and interpolating to get voxelwise displacements

9. Scale the image by a factor $s \sim \mathcal{U}(0.4,\ 1.6)$ and crop or pad the resulting image (using a linear ramping strategy) to preserve image shape

During training of the deep learning model, the above steps are performed in a stacked manner (i.e. after one another). For every image, the processing functions above are each applied with a probability p (chosen to be 0.5 as in the original paper). The intervals specifying the parameter distributions indicated above was chosen to be the same as in [15], and should together with the probabilities be interpreted as hyperparameters of the model which can be optimized. In addition to the augmentation steps given above, another step was added which was not in the original paper, namely changing a number of slices in the top and bottom of the images to a set value - either the maximum value of the

image or the minimum value, with equal probabilities. This was added to compensate for the padding of a number of images (see the discussion in connection to Figure 7), and the number of slices affected was chosen randomly so that 0-15% of the voxels were changed, in accordance with Figure 7. Just like the other augmentation steps, this addition had a probability of 0.5 to be utilized for each batch during the training of the segmentation network.

It should be noted that this approach to domain adaptation comes with a big practical advantage, namely that it does not require any specific knowledge regarding the *target* domain, as compared to the two methods described below which both require access to unlabeled image data.

4.2.2 Weak Labels

As stated previously, manually delineating MR images is a time-consuming and expensive task (this is why we are interested in automatic segmentation in the first place). However, there are other ways of automatically segmenting MR images than the deep learning approach discussed in this report. Such automatically generated labels will generally not be of as high quality as labels delineated by experts, but it can still be expected that automatically generated labels are correlated to the ground truth. This raises the question as to whether automatically generated labels can be used for training a deep learning model, that will in turn exceed the performance of the algorithm that generated the labels (on some held out validation set). While this might seem unintuitive (after all, how could the student outperform the teacher?) training with noisy labels is an important subfield within the study of weakly-supervised learning [16]. What is more; if the label noise is unbiased one might expect that the noise cancels out, so that the expected value of the noisy labels coincide with the exact labels. An excellent example of this is given by Rolnick et. al. [17] who trained a classification network on the MNIST data set. With label accuracy just 1% higher than for randomly assigned labels, they still managed to achive 90% accuracy on a held-out test set with noise-free labels.

The success of using noisy labels as augmentation in a one-domain setting sparks the question if the same technique is successful for reducing the domain shift error, and motivates the experiments described in this section.

A way to create weak labels for the target domain images is to use a technique called atlas-based segmentation, which was the state-of-the-art technique for medical image segmentation before the deep learning paradigm shift. This technique is based on deformable image registration, where the RoI:s from one or a number of labeled patients (atlases) are deformed to match the geometry of the image one wants to generate segmentations for. The resulting deformation is then used to deform the labels of the atlas, generating the segmentation. For multi-atlas segmentation, the atlases are deformed one-by-one, and the result of each of these is fused using one of a number of different possible algorithms [18].

In the multi-domain setting described in this report, labels were created for the 52 unlabeled samples of Target Domain 2: $\mathcal{D}_{T2}^{unlabeled} = \{\mathbf{x}_i^{T2}\}_{i=1}^{52}$. They were generated using an implementation of multi-atlas segmentation in RayStation, and the deep learning segmentation network was then trained on the combined set $\mathcal{D}_S^{train} \cup \{(\mathbf{x}_i^{T2}, \mathbf{y}_i^{*T2})\}_{i=1}^{52}$ where

$\mathbf{y}_i^{*T2}$ denotes the weak label of sample i.

Examples of such images with their corresponding weak labels are shown in Figure 9. Generally, the atlas-segmented labels succeed in localizing where the RoI:s are in the image. However, they are not good at making precise predictions regarding the boundaries of the RoI:s, and often miss these (although the RoI boundaries are correctly located in some of the cases).

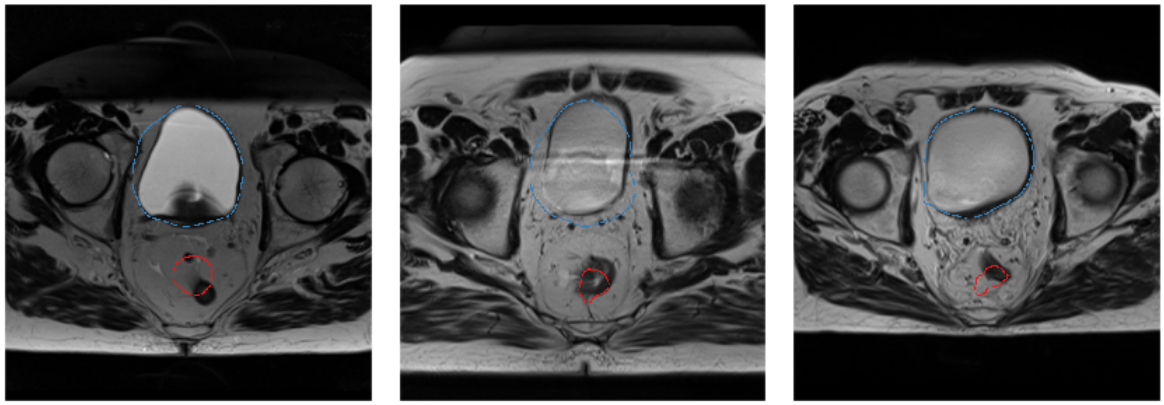


Figure 9: Samples of $\mathcal{D}_{T2}$ images with weak labels. The weak labels generally do an adequate job of delineating the RoI:s, however they are not very precise and often miss the boundaries.

4.2.3 Cycle GAN domain transfer

Another popular method for reducing the domain shift error is to align the images of the source and target domains, so that they look more similar [7]. There exists many techniques for this type of domain transfer, but one of the prominent ones is to use a Generative Adversarial Network (GAN) [19] to synthesize new training data. For the specific problem of domain shift, utilising a Cycle GAN [20] for the domain translation is particularly interesting, as it allows mapping images from one domain to another without losing geometrical information (i.e. RoI placement), allowing style changes in the source images s.t. they resemble target images, while still staying true to their labels.

The Cycle GAN consists of two parallel data flows where the principal idea is that two generators are trained, one for the mapping from source to target domains

For the training of the Cycle GAN in the following experiments, images from $\mathcal{D}_S^{train}$ and the 52 unlabeled images from $\mathcal{D}_{T2}$ were used to train the Cycle GAN to learn the mapping between the source domain and target domain 1. Therafter, the Source-to-Target Generator G_T was used to generate synthetic data

$$\mathcal{D}_{T2}^{synthetic} = \{G_T(\mathbf{x}_i^S), \mathbf{y}_i^S\}_{i=1}^{n_S^{train}} \tag{9}$$

i.e. the labels of the $\mathcal{D}_S$ training data were used as labels for the synthetic data set. The deep learning model was then trained on the combined set $\mathcal{D}_S^{train} \cup \mathcal{D}_{T2}^{synthetic}$. For each

epoch in the training, all samples of this dataset were used, meaning that the model trains on two versions of the same image per epoch; the original image and the domain-mapped image. Because of this, the number of epochs were reduced by half as compared to the other experiments in this report.

4.2.4 Additional methods

The weak labels and Cycle GAN approaches to data augmentation both have their drawbacks. The obvious downside with the weak labels approach is that the weak labels are not perfect, as seen in Figure 9. For the Cycle GAN, even if target domain specific image features such as intensity levels, noise and artefacts are reproduced in the synthetic data, geometrical features such as RoI shape are not. A domain transfer method that remedies both of these drawbacks have been shown to achieve good results on brain MRI data [1], and was tested for this thesis as well. The method builds on a deep learning based image registration method where the source domain geometries are deformed to match unlabeled target domain data. However, it proved difficult to transfer these results to the pelvic region, possibly because the image registration problem is harder to solve for this region. Therefore, the method was never used in the data augmentation setting. The method is described in Appendix A, together with some initial results.

4.3 Evaluation

Three different metrics are used to evaluate the performance of the segmentation network. Defining V as the ground truth 3D mask of a specific class, and $\hat{V}$ the corresponding predicted mask, these are:

- **DICE.** The Dice score measures the volumetric overlap of the predicted label mask and the true mask, by calculating

$$\frac{2|V \cap \hat{V}|}{|V| + |\hat{V}|} \tag{10}$$

  where $|V|$ denotes the number of voxels of V

- **Hausdorff distance** The Hausdorff distance measures the maximum shortest distance between a surface element of V_c and one of $\hat{V}c$, i.e.

$$H_{max} = \max\left\{ \sup_{\hat{v}\in\partial\hat{V}} \left(\inf_{v\in\partial V} (d(v,\, \hat{v})) \right),\ \sup_{v\in\partial V} \left(\inf_{\hat{v}\in\partial\hat{V}} (d(v,\, \hat{v})) \right) \right\} \tag{11}$$

  where d is the Euclidean distance and ∂V and $\partial\hat{V}$ denote the surfaces of V and $\hat{V}$, respectively. Similarly, instead of taking the maximum distance (the supremum norm) one can calculate some percentile, creating for example the $H_{95\%}$ distance. Figure 10 shows the visual interpretation of the components used for calculating the distance.

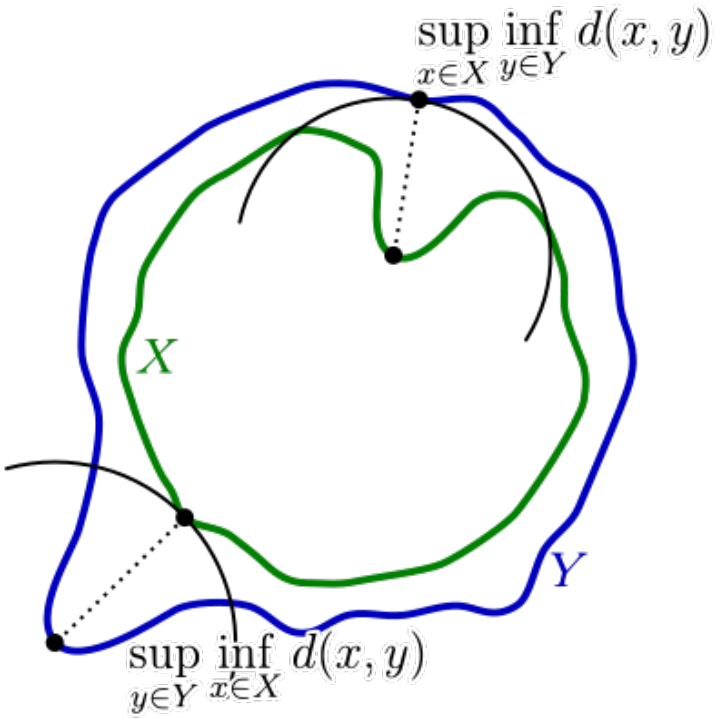


Figure 10: Visualization of the two elements used when calculating the H_{max} distance (see Equation (11)). Reproduced from Rocchini, CC BY 3.0, via Wikimedia Commons.

A note on the calculation of results for the rectum RoI, is that the ground truth contours vary in how much of the RoI is delineated. This might be because of varying guidelines or interpretations of these, regarding how much of the sigmoid colon and anal canal to include in the delineation of the rectum. The results of this is that there is variation between and within the different domains regarding what is included in the ground truth rectum contours. To mitigate the effect this has on the performance metrics, the rectum RoI:s were truncated so that no part of the RoI in the model predictions extends above or below the ground truth, and vice versa.

The experiments are evaluated on the means and standard deviations of the defined metrics, evaluated over all patients. Additionally, the score distributions are evaluated, and the model predictions are manually inspected qualitatively. As each model was trained more than once, all patients occur multiple times in the evaluations. This means that if a model was retrained 5 times, for example, all of the patients occur 5 times in the evaluation.

5 Results

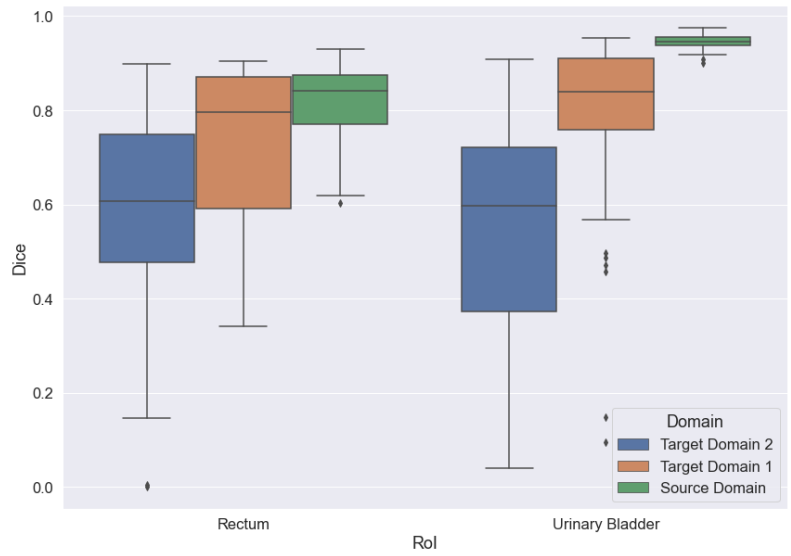
5.1 Baseline segmentation results

Figure 11 shows the score distributions for the baseline model after repeating training 7 times. The domain shift error is apparent for all metrics, both in terms of the absolute scores and the variability of the scores. Noticeably, there is not only a performance difference between the source and target domains, but there is also a large gap between $\mathcal{D}_{T1}$ and $\mathcal{D}_{T2}$ except for on the $H_{95\%}$ scores for the rectum. This indicates that the distance between the probability distributions from which the samples are obtained is smaller between $\mathcal{D}_S$ and $\mathcal{D}_{T1}$ than between $\mathcal{D}_S$ and $\mathcal{D}_{T2}$, something that is also indicated in the image statistics of Figures 4, 5 and 6, i.e. the RoI volumes and mean intensities as well as the first principal components of the normalized image histograms.

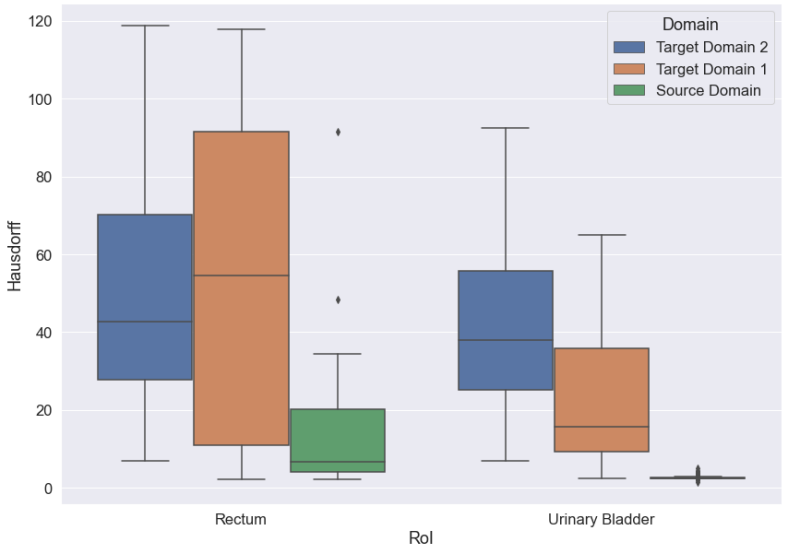
Upon inspection of the predicted segmentation masks (see examples in Figure 12) it is clear that while the segmentation model makes mistakes on the source domain (e.g.bottom row, left column in Figure 12), these are not as frequently occuring nor as large as for the target domains - especially $\mathcal{D}_{T2}$. From Figure 12, two distinct types of prediction errors are illustrated:

- prediction follows ground truth quite nicely for part of the contour, but makes some large mistake by either 'floating out' (e.g. top row, middle column) or 'collapsing in' (e.g. top row, rigt column);
- the model predicts ROIs elsewhere in the image, seemingly random (e.g. middle row, right column).

While both types of errors occur in all domains, they are significantly more common in the target domains. Especially the second type of error is evident in almost all examples from $\mathcal{D}_{T2}$, and likely affects the $H_{95\%}$ distance significantly. Note that both types of error contribute to making the predicted RoI masks unrealistic in shape, i.e. it is enough to observe the predicted masks to realize that they are wrong, without even seeing the corresponding MR image.



(a) Dice score



(b) Hausdorff 95 score

Figure 11: Results for the baseline experiments, evaluated over the different domains. The domain-dependent performance difference is clear, in terms of all metrics and RoI:s. Note escpecially that the worst-performing DICE scores on $\mathcal{D}_{T2}$ are lower than 0.1, and that a significant number of samples in both target domains have a $H_{95\%}$ of more than 5 cm for the rectum, which is far from clinically acceptable.

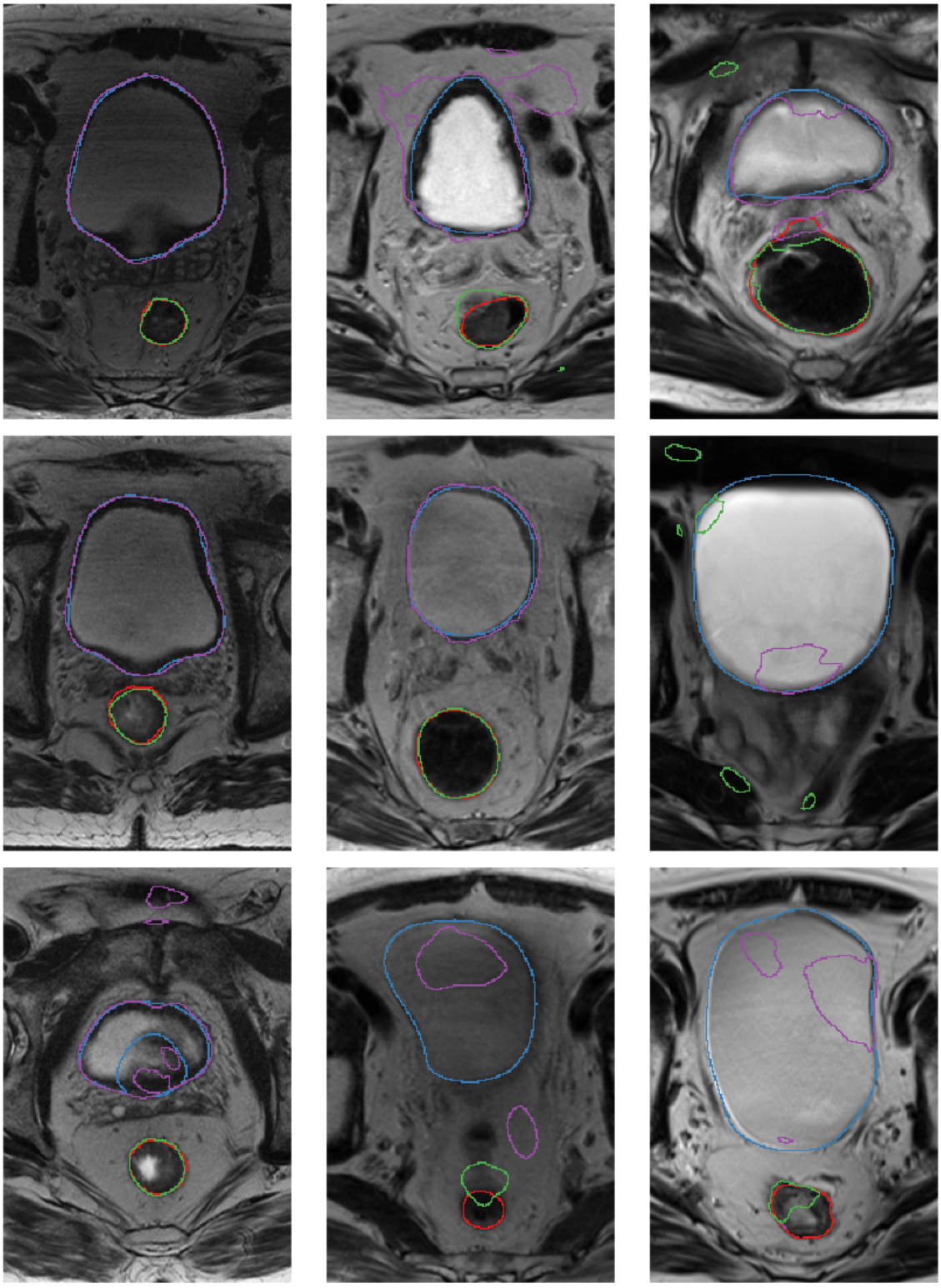


Figure 12: Sample of the segmentation results for the baseline model on 9 patients, with columns corresponding to $\mathcal{D}_S$ (left), $\mathcal{D}_{T1}$ (middle) and $\mathcal{D}_{T2}$ (right). Ground truth: rectum = red, bladder = blue. Predictions: rectum = green, bladder = purple. For $\mathcal{D}_S$, the predictions follow the ground truth quite nicely, although some errors are made. For the target domains however, the number of errors is greatly increased - especially for $\mathcal{D}_{T2}$. There are at least two distinct types of prediction error; firstly there are instances where connection to the ground truth is lost, secondly there are instances where the model predicts RoI structures seemingly random in the image.

Table 2: Average difference in performance metrics for the urinary bladder between **a)** including a specific image processing step in the BigAug pipeline and **b)** not including it. The three top scores for every metric is bolded. Note that negative differences in the $H_{95\%}$ distances indicate better performance.

	$\mathcal{D}_S$		$\mathcal{D}_{T1}$		$\mathcal{D}_{T2}$	
	Dice [%]	$H_{95\%}$ [mm]	Dice [%]	$H_{95\%}$ [mm]	Dice [%]	$H_{95\%}$ [mm]
rotation	**0**	**-0.7**	**4**	**-9.8**	**10**	**-9.3**
additive pixelwise noise	**0**	0.1	0	0.3	-1	0.5
elastic	**0**	**-0.8**	2	-4.7	**7**	**-5.4**
blurring	**0**	0.3	-1	-0.3	2	-3.2
sharpening	**0**	**-0.6**	**6**	-4.7	**7**	-1.1
intensity shift	**0**	-0.3	2	-1.5	0	-0.3
gamma correction	**0**	-0.1	3	**-7.6**	6	-5.3
affine intensity transformation	**0**	0.3	-3	3.1	-5	2.2
scaling	**0**	-0.2	**8**	**-9.0**	**12**	**-7.5**

5.2 Choice of BigAug augmentation steps

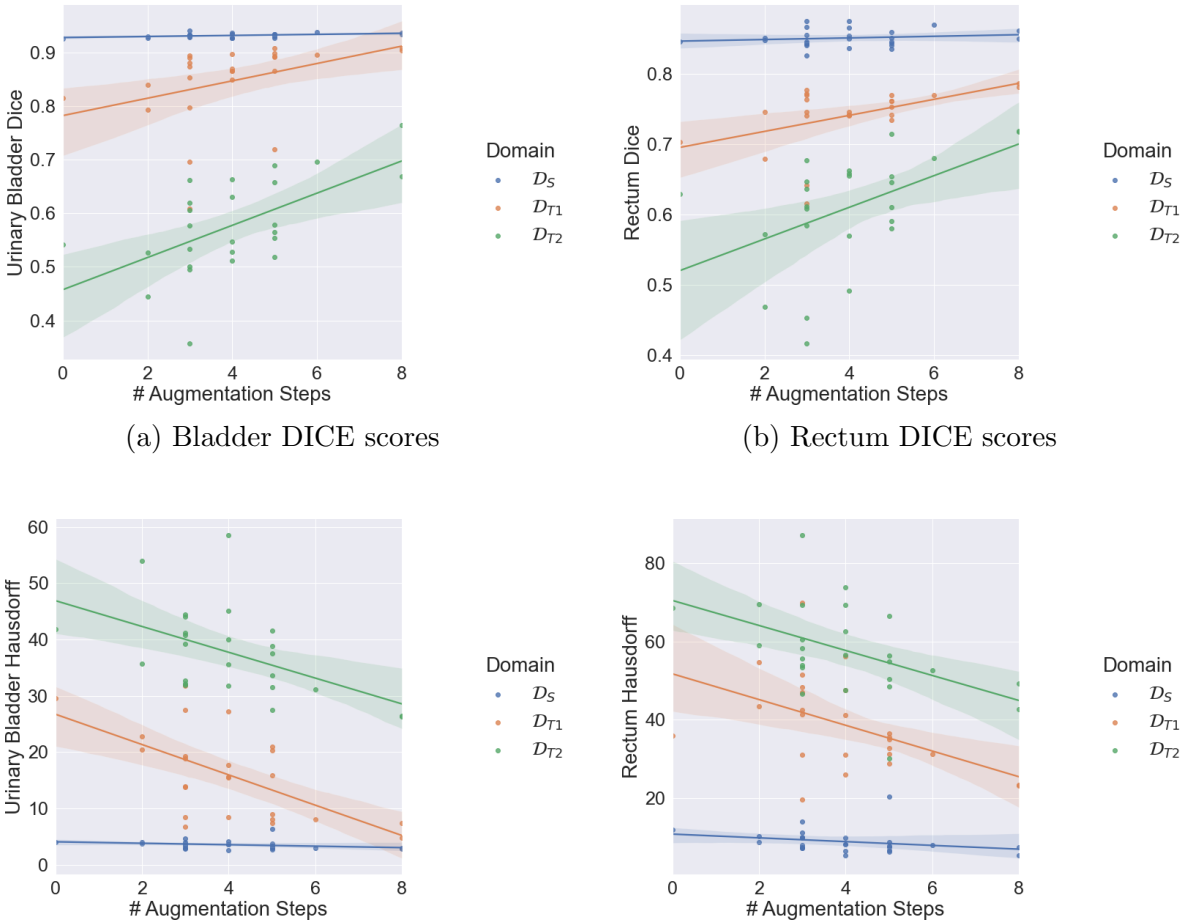
In order to get a better understanding of how the 9 separate steps of the BigAug pipeline affects model performance, a hyperparameter search was carried out. The search consisted of repeatedly training a segmentation model using the BigAug pipeline, but before each new training session, all 9 steps were randomly chosen to be either used or not used for the training, with a probability of 50%. To keep training times low, a stop criterion was included so that training was aborted if no improvement on the validation loss was registered for 100 consecutive epochs. This process was repeated 25 times, and the evaluation was done on the validation sets of each domain. The added augmentation step of masking the images to simulate padding was not used in any of these, and the metrics were calculated on the full rectum RoI:s, not the truncated ones.

For each of the 9 image processing steps in BigAug, the average metrics were compared between the experiments where the step was included and the ones where it was not. The results of these experiments are shown in Tables 2 and 3 for the urinary bladder and rectum, respectively. As is clear from the tables, all of the preprocessing steps except for the additive noise and intensity shift (steps 3 and 4 as described in Section 4.2.1) were in the top 3 steps with highest impact for some metric on at least one of the target domains. No proper statistical investigation of these results was carried through, as it was deemed that the number of experiments was far too small to take interactions between augmentation steps into consideration. Instead, the conclusions from these results were that all of the augmentation steps might have an impact on segmentation results and that none of the steps should therefore be excluded from the pipeline. Note that in general, the impact of the augmentation steps on $\mathcal{D}_S$ results is significantly smaller than on $\mathcal{D}_{T1}$ and $\mathcal{D}_{T2}$. An observation regarding the results in Tables 2 and 3 is that the two augmentation steps that seem to have the biggest performance impact are *rotation* and *scaling*. These two steps, however, may both invoke the need for padding in the resulting images. This means that it is not certain that it was the actual rotation and scaling of the images that affected the results, but that it could also be the resulting padding that is boosted performance.

Table 3: Average difference in performance metrics for the rectum between **a)** including a specific image processing step in the BigAug pipeline and **b)** not including it. The three top scores for every metric is bolded. Note that negative differences in the $H_{95\%}$ distances indicate better performance.

	$\mathcal{D}_S$		$\mathcal{D}_{T1}$		$\mathcal{D}_{T2}$	
	Dice [%]	$H_{95\%}$ [mm]	Dice [%]	$H_{95\%}$ [mm]	Dice [%]	$H_{95\%}$ [mm]
rotation	**0**	-0.6	**5**	**-12.6**	**10**	**-12.6**
additive pixelwise noise	**0**	-1.1	0	-0.2	-1	-2.7
elastic	**0**	1.9	**3**	-1.8	**6**	1.4
blurring	-1	0.2	1	-5.7	4	**-10.0**
sharpening	**1**	0.6	2	1.6	3	6
intensity shift	**1**	**-1.2**	1	0.1	0	-2.3
gamma correction	**0**	**-2.6**	1	**-7.1**	3	-5
affine intensity transformation	**0**	1.5	-1	**-7.2**	-5	-1.4
scaling	**0**	**-3.7**	**3**	-6.8	**9**	**-10.7**

In addition to the results in Tables 2 and 3, the experiments were used to study the effect of varying the number of augmentation steps for model training. The results are shown in Figure 13, and clearly indicate a correlation between model performance and the number of augmentation steps. This further motivates that all of the augmentation steps were used for the final results and evaluation on the test sets.



(a) Bladder DICE scores

(b) Rectum DICE scores

(c) Bladder Hausdorff 95 scores

(d) Rectum Hausdorff 95 scores

Figure 13: DICE score and $H_{95\%}$ distances plotted against number of augmentation steps in the BigAug pipeline, together with least squares-fitted regression lines (including confidence intervals). It is clear that there is a connection between model performance and number of augmentation steps for the two target domains. For the source domain, there is no clear correlation.

5.3 Training of Cycle GAN for domain transfer

During the training of the Cycle GAN, tendencies of mode collapse in the creation of synthetic target data was observed. In many of the samples, the target generator G_T synthesizes a bright bladder in the bowel region of the patient. It is understandable that this feature might trick the adversarial discriminator, as very bright bladders are common in $\mathcal{D}_{T2}$ but not in $\mathcal{D}_S$. The cycle consistency loss should incentivise the conservation of geometrical features in the generation of synthetic data in order to be able to generate a good cycled image, but apparently this is not the case. What is more, the cycled image (the result of feeding $G_T(\mathbf{x}_i^S)$ through the target-to-source generator G_S) is nearly identical to the original image - see Figure 14 below.

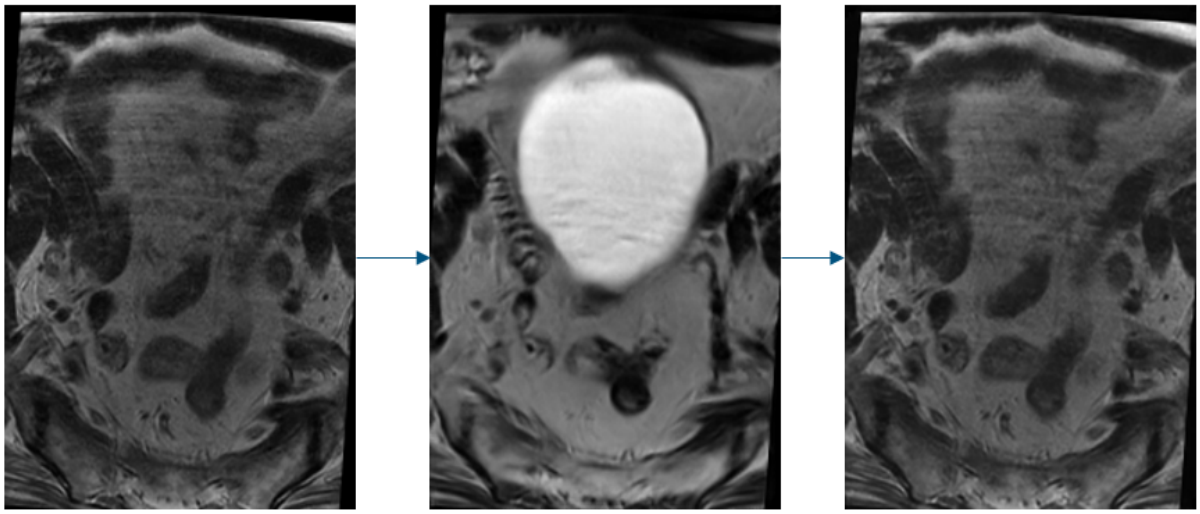


Figure 14: From left to right: original image, synthetic target domain image and cycled. The cycle gan introduces new geometrical features but still manages to get nearly perfect cycled images, indicating that some information about the original image (left) is hidden in the generated image (middle). This behavior enables low cycle-consistency loss even though new geometric features are introduced in the generation process.

The behavior described above has been documented with Cycle GANs previously [21], and the explanation is that the generator is hiding information about the original image that is not visible to the naked eye, enabling perfect restoration.

As the hidden information is not visible to humans, it is reasonable to assume that it is hidden in low amplitude (possible high frequency) intensities in the generated image, which would indicate that adding low-amplitude noise to the synthetic data would make perfect reproductions of the original image impossible. Figure 15 indicates that this is the case, as it shows that a very small perturbation of $G_T(\mathbf{x}^S)$ essentially breaks down the cycled image.

Because of the behavior discussed above, the Cycle GAN was re-trained using a slightly different pipeline, where Gaussian noise was added to the generated data before they are passed through the second generator. This resulted in a more robust model with smaller tendency to create artificial bladders in the synthetic data, although not completely making this behavior disappear. Importantly, this new Cycle GAN was not able to perfectly reproduce the initial images, meaning that adding geometrical information that is not

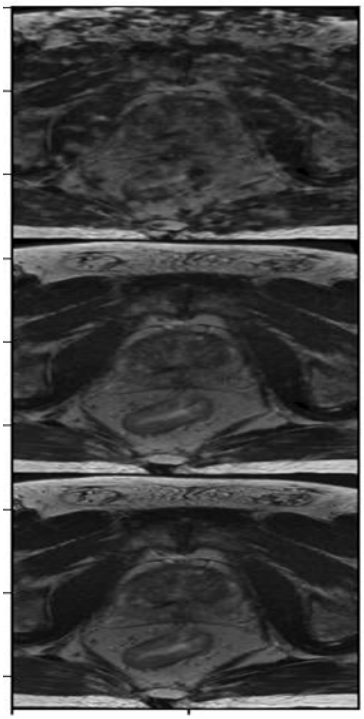


Figure 15: Cycled image after perturbation (top), without perturbation (middle) and original image (bottom). Adding noise on a scale that is inconceivable to the naked eye makes the recreation of the original image break down, as apparent in the differences between the top and middle images.

present in the original image leads to an increase in the cycle consistency loss. See Figure 16 for an example of this. It is likely that the unwanted behavior of adding artificial bladders can be reduced be reweighting the adversarial and cycle consistency loss functions, however this was not tested due to time constraints.

It should be noted here that while the Cycle GAN generates synthetic data that seem close to the distribution of $\mathcal{D}_{T2}$, the discriminators clearly overfits to the training data - see training curves in Appendix B. This might partially explain the mode collapse behavior described above, as it means that in order to fool the discriminators, the generators have to synthesize data that looks just like the training data from the other domain (not just as if it is drawn from the same distribution).

5.3.1 Cycle GAN generated data

After training the cycle GAN for 100 epochs, the model of epoch number 50 was chosen to generate synthetic training data for the segmentation model. Figure 17 shows one example of synthetic training data generated by this model in the source - target direction and one example in the target - source direction. Both images are from the training data of the Cycle GAN, as this is what is used for the subsequent training of the segmentation model. It is apparent that the Cycle GAN manages to modify images so that they resemble the other domain, such as the intensity variations in the bladder and shadow-like artefact of $\mathcal{D}_{T2}$, as well as the details in the fat of $\mathcal{D}_{\mathcal{S}}$.

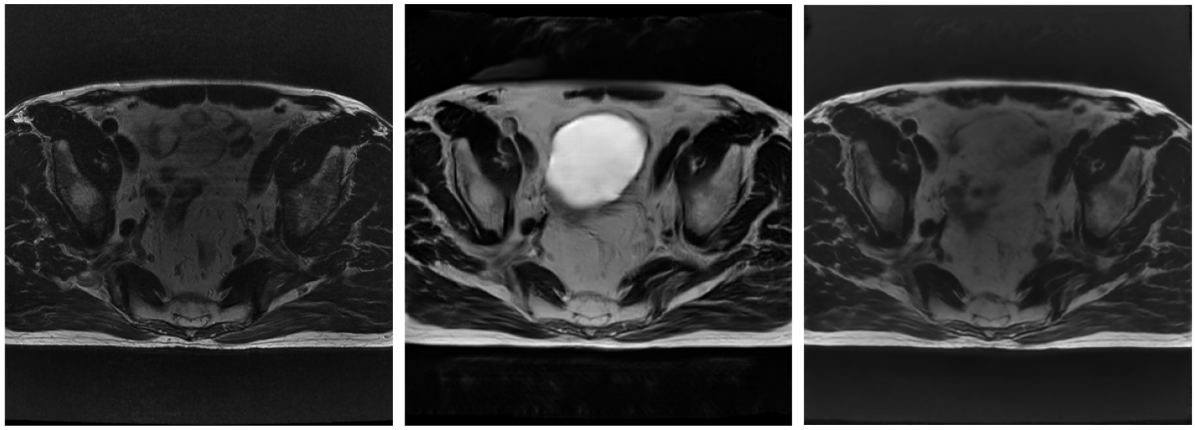


Figure 16: Left to right: original image, synthetic target domain image and cycled original image, after adding intermediary Gaussian noise to the training pipeline. The generator still synthesizes new geometrical information (the bladder in the synthetic image), but this now increases cycle inconsistency, as indicated by the differences between the original image and the cycled image.

In order to study the nature of the synthetic data, the same statistics were calculated as for the real images (see Figures 5 and 6). The results of this is shown in Figures 19 and 18. These figures, together with the samples in 17, indicate that the distributions of $\mathcal{D}_{T2}$ and $\mathcal{D}_{T2}^{synthetic}$ are closer to each other than to $\mathcal{D}_S$, meaning that the Cycle GAN serves its purpose as it can generate synthetic data from a distribution that resembles that of the target domain distribution.

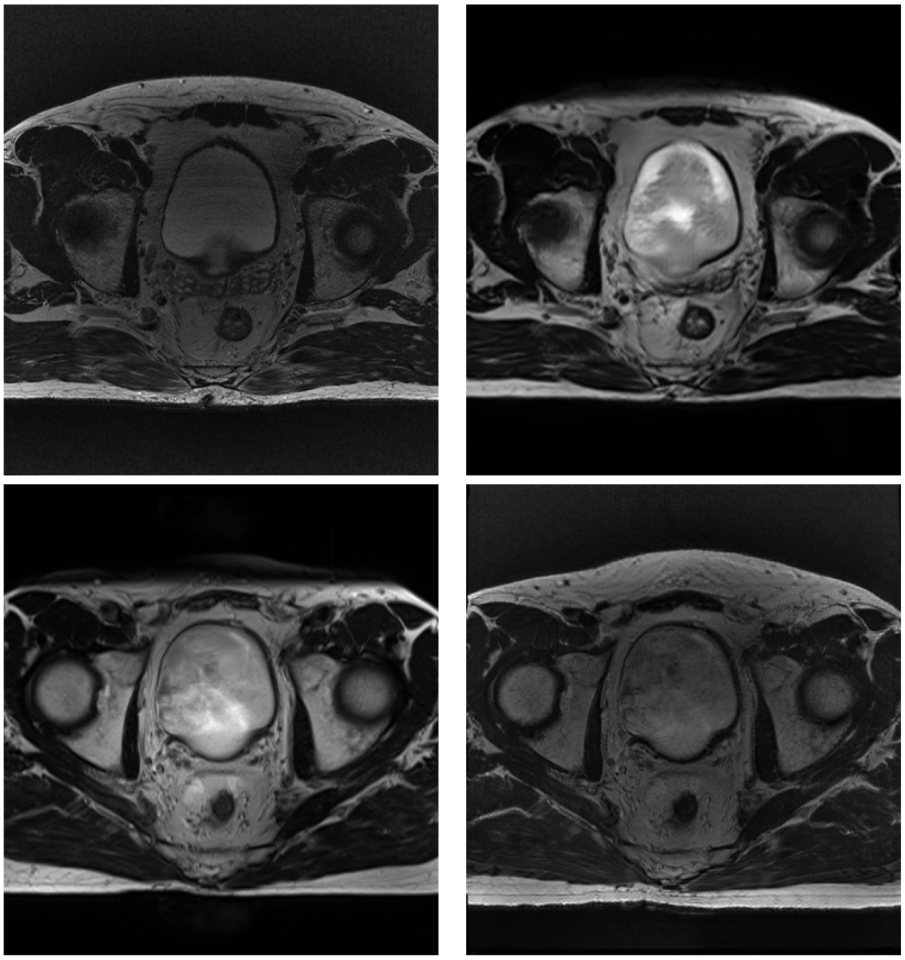


Figure 17: Samples of domain transfer data generated by the Cycle GAN. In the top row, a sample from $\mathcal{D}_S$ (left) and the the synthetic $\mathcal{D}_{T2}$ data generated from it. Vice versa in the bottom row, i.e. a $\mathcal{D}_{T2}$ image and its synthetic $\mathcal{D}_S$ counterpart. Domain-specific intensity distributions and details are picked up by the Cycle GAN generators to produce realistic synthetic data

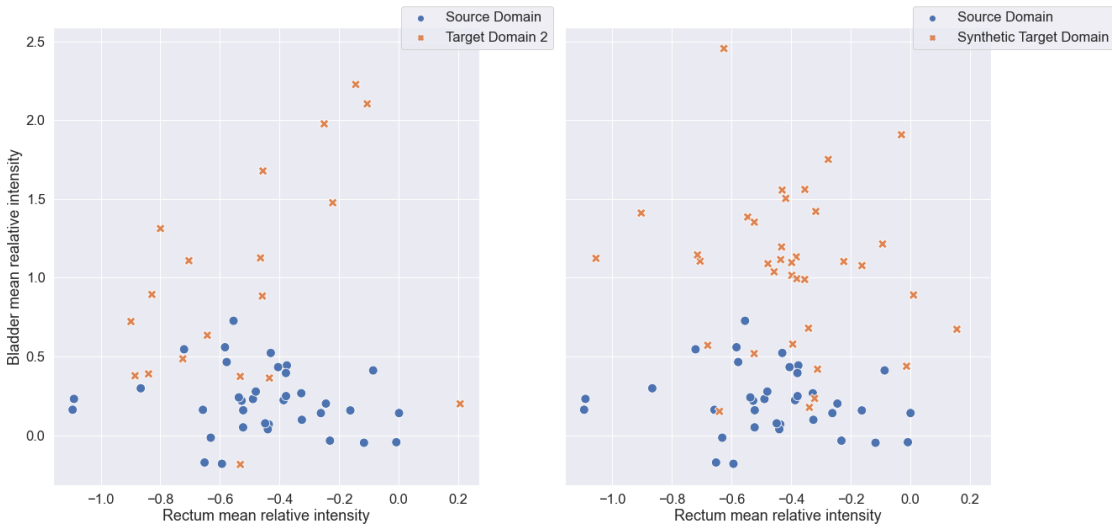


Figure 18: RoI Mean intensities. To the left is the real data from $\mathcal{D}_S$ and $\mathcal{D}_{T2}$. To the right, the $\mathcal{D}_{T2}$ data are replaced by synthetic data generated by running the $\mathcal{D}_S$ data through the Cycle GAN generator. The bright bladders of $\mathcal{D}_{T2}$ are reproduced in the synthetic data, and the relative intensity distributions of the rectum are not changed significantly.

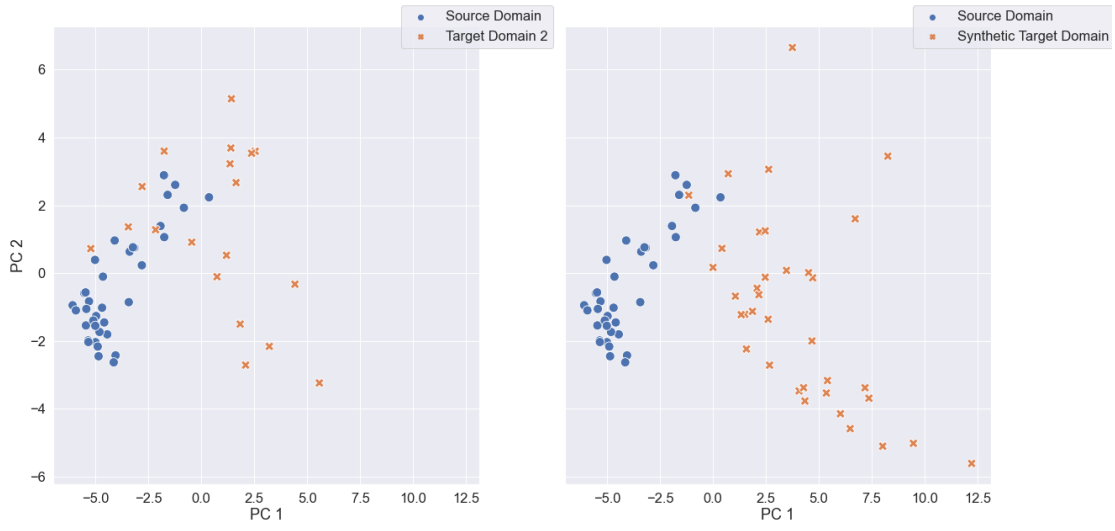


Figure 19: First two principal components of the histograms of normalized images from $\mathcal{D}_S$ and $\mathcal{D}_{T2}$ (left), as well as $\mathcal{D}_S$ and synthetic data to imitate the style of $\mathcal{D}_{T2}$ (right). The synthetic images seem to follow the intensity distribution of $\mathcal{D}_{T2}$ quite well. The principal components differ from those of Figure 6 as the synthetic data is included in the calculation of the principal components.

Table 4: Performance comparison for all experiments on $\mathcal{D}_S$. Best scores in bold and top 3 (excluding best) in italic. Overall, Cycle GAN + BigAug give the best results.

	Rectum		Urinary Bladder	
	Dice [%]	$H_{95\%}$ [mm]	Dice [%]	$H_{95\%}$ [mm]
Baseline	82 (8)	13.4 (14.7)	**95** (2)	*2.7* (0.6)
Cycle GAN	**84** (8)	9.7 (9.0)	**95** (1)	**2.6** (0.6)
BigAug	83 (8)	9.9 (7.9)	94 (3)	4.2 (7.2)
Weak Labels	83 (8)	9.8 (8.1)	94 (2)	3.1 (1.3)
Weak Labels + BigAug	83 (8)	*8.1* (6.2)	94 (3)	3.4 (2.3)
Cycle GAN + BigAug	**84** (6)	*7.9* (6.0)	**95** (2)	*2.7* (0.9)
Cycle GAN + Weak Labels	83 (7)	10.4 (9.0)	94 (2)	3.4 (1.7)
Cycle GAN + Weak Labels + BigAug	**84** (6)	**7.8** (5.4)	94 (3)	3.4 (2.1)

5.4 Quantitative evaluation of augmentation methods

For each of the augmentation methods discussed in Section 4.2, the segmentation network was trained 5 times. This section covers these results for each of the domains studied. Boxplots visualizing the distributions of the performance metrics defined in Section 4.3 for all of the experiments are given in Figures 20 and 21 (for the urinary bladder and rectum, respectively), and Tables 4, 5 and 6 give the means and standard deviations of the metrics for each domain respectively. These tables and figures show the results of evaluating model performance on the held out test sets on each of the domain. Since optimization on the target domain validation sets was kept low, the results on these can also be informative, and are displayed in appendix C.

For $\mathcal{D}_S$, it is clear from Table 4 that the average metrics for the rectum were improved in all of the experiments, with the best overall results obtained by combining the use of Cycle GAN-generated data with the BigAug pipeline. The boxplots in Figure 21 show that the performance metrics are improved especially for the patients on which they perform poorly - the worst $H_{95\%}$ distance for the baseline is > 90 mm, while it is < 20 mm for the combination of all three methods (see Figure 21b). For the urinary bladder, no method outperforms the baseline in terms of DICE score, but the $H_{95\%}$ distances are slightly improved for the 'Cycle GAN' and 'Cycle GAN + BigAug' experiments. Note especially that all experiments where weak labels were utilised led to an increase in the average $H_{95\%}$ distances. In the boxplots of Figure 20b, this is seen by the fact that the outliers have larger $H_{95\%}$ distances for the Weak Labels experiments.

For $\mathcal{D}_{T1}$, there is a clear improvement from the baseline in all of the experiments, as seen in Table 5 as well as Figures 20 and 21. For many of the experiments, Figure 21 indicate that there is no clear difference in the results between $\mathcal{D}_{T1}$ and $\mathcal{D}_S$ patients, suggesting that the domain gap is effectively bridged. For the urinary bladder, the performance loss compared to the source domain is still visible, however the differences are significantly smaller than for the baseline. Furthermore, from Table 5 it is clear that combining methods lead to better average performance (except for the Cycle GAN + Weak Labels combination). The best average result is obtained by the Cycle GAN + BigAug combination, although the differences to the results of combining all three methods are

Table 5: Performance comparison for all experiments on $\mathcal{D}_{T1}$, as well as scores for 5 human experts compared to their consensus. Best model scores in bold, and top 3 (excluding best) in italic. Cycle GAN + BigAug gives the best results for all metrics.

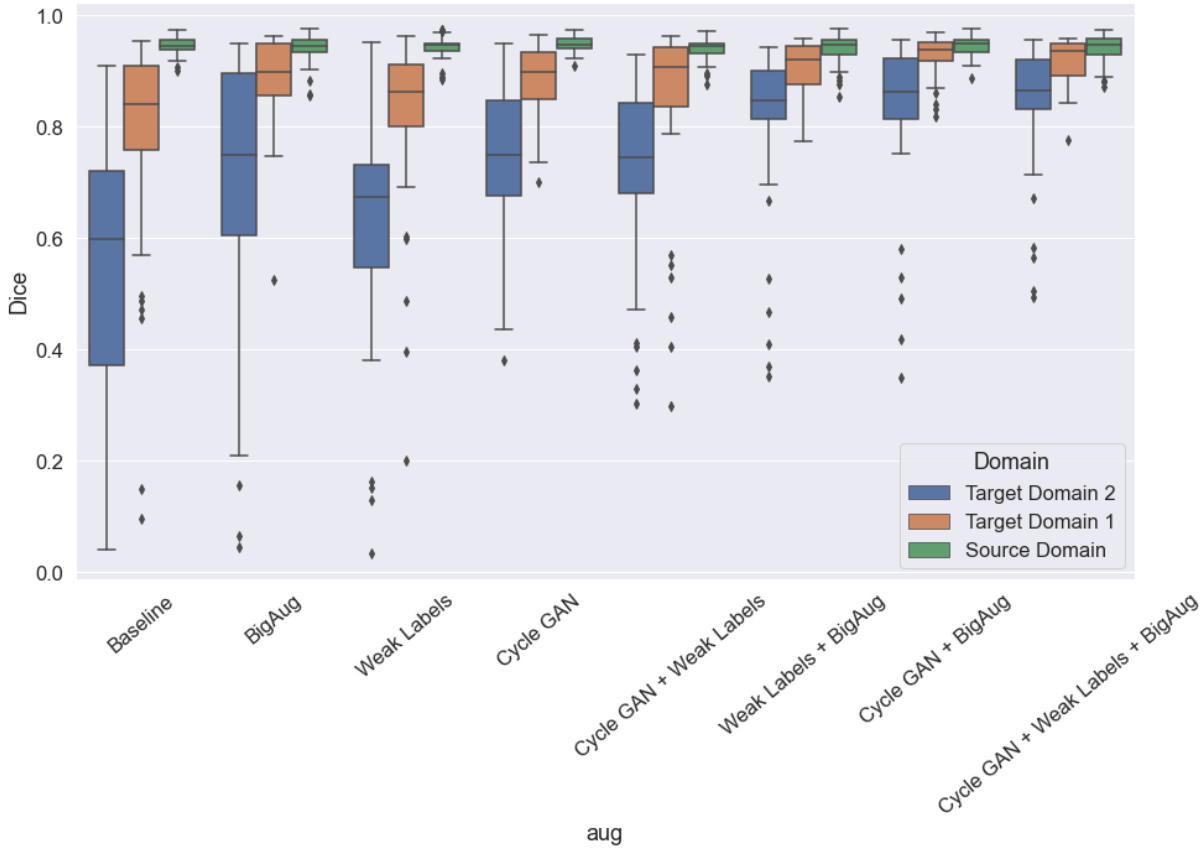
	Rectum		Urinary Bladder	
	Dice [%]	$H_{95\%}$ [mm]	Dice [%]	$H_{95\%}$ [mm]
Baseline	73 (16)	53.2 (40.2)	79 (17)	23.6 (19.0)
Cycle GAN	*85* (7)	20.5 (28.3)	89 (6)	9.3 (7.9)
BigAug	83 (9)	10.4 (6.6)	89 (8)	12.1 (18.5)
Weak Labels	84 (7)	10.5 (15.7)	82 (15)	20.1 (21.9)
Weak Labels + BigAug	84 (5)	*8.2* (4.5)	*91* (4)	*6.5* (5.9)
Cycle GAN + BigAug	**87** (5)	**7.9** (5.6)	**92** (4)	**4.3** (4.3)
Cycle GAN + Weak Labels	78 (10)	16.7 (24.9)	84 (16)	12.7 (15.9)
Cycle GAN + Weak Labels + BigAug	*86* (4)	*9.0* (5.8)	**92** (4)	*4.5* (2.7)
Interobserver agreement	93 (2)	3.1 (5.3)	95 (2)	2.4 (0.7)

quite small - especially for the urinary bladder. As stated in Section 3.2, the ground truth for domain $\mathcal{D}_{T1}$ was obtained by 5 clinical practitioners agreeing on a consensus delineation. Table 5 also shows the average metrics obtained by comparing the individual delineations of each of the experts to this consensus. It is clear that none of the models reach the same level of average performance as these 5 domain experts, however it should be noted that for some of the patients the model performs on par with the practitioners.

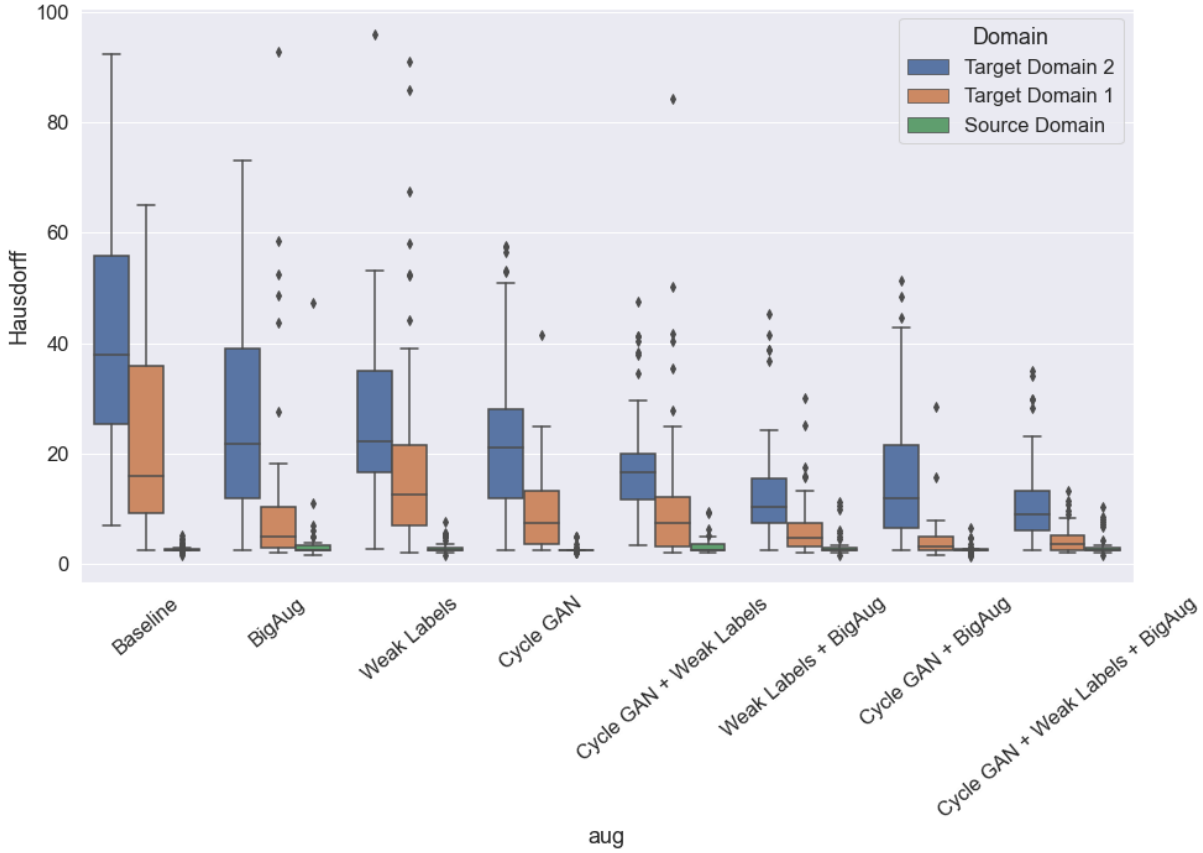
For $\mathcal{D}_{T2}$ too, all methods provide significant improvement as compared to the baseline, as shown in Table 6 and Figures 20-21. When only one of the methods is used, the introduction of synthetic data (Cycle GAN experiments) or weak labels results in higher performance on the rectum than the BigAug pipeline, in contrast to the results for $\mathcal{D}_{T1}$ where this is not the case. For the urinary bladder, the Cycle GAN experiments show the best results. It is clear that the combination of methods additionally boosts the segmentation performance (except for the DICE scores of Cycle GAN + Weak Labels), and the best average results are obtained when all three methods are combined. For the rectum, Figure 21 indicate no clear differences in performance compared to the source domain for these experiments (for the bladder, the results still lag significantly behind the other two domains). In contrast to the results on the $\mathcal{D}_S$ and $\mathcal{D}_{T1}$, the inclusion of weakly labeled data in general leads to clear performance gains on the $H_{95\%}$ distances for $\mathcal{D}_{T2}$(compare for example 'BigAug' with 'Weak Labels + BigAug' in Figure 21b).

Table 6: Performance comparison for all experiments on $\mathcal{D}_{T2}$. Best model scores in bold, and top 3 in italic. In general, the combination of all three augmentation methods give the best results.

	Rectum		Urinary Bladder	
	Dice [%]	$H_{95\%}$ [mm]	Dice [%]	$H_{95\%}$ [mm]
Baseline	57 (22)	49.9 (28.9)	54 (24)	40.3 (19.6)
Cycle GAN	80 (9)	17.9 (20.5)	75 (14)	23.3 (15.9)
BigAug	72 (13)	34.7 (31.3)	70 (23)	25.6 (19.0)
Weak Labels	75 (10)	16.5 (18.4)	64 (22)	25.0 (16.8)
Weak Labels + BigAug	*81* (8)	*7.9* (4.7)	*81* (14)	*13.4* (10.3)
Cycle GAN + BigAug	**83** (8)	11.2 (16.5)	*83* (13)	*15.7* (13.3)
Cycle GAN + Weak Labels	68 (12)	*10.8* (5.2)	73 (16)	18.1 (10.7)
Cycle GAN + Weak Labels + BigAug	*82* (7)	**7.7** (5.7)	**84** (11)	**11.2** (8.2)

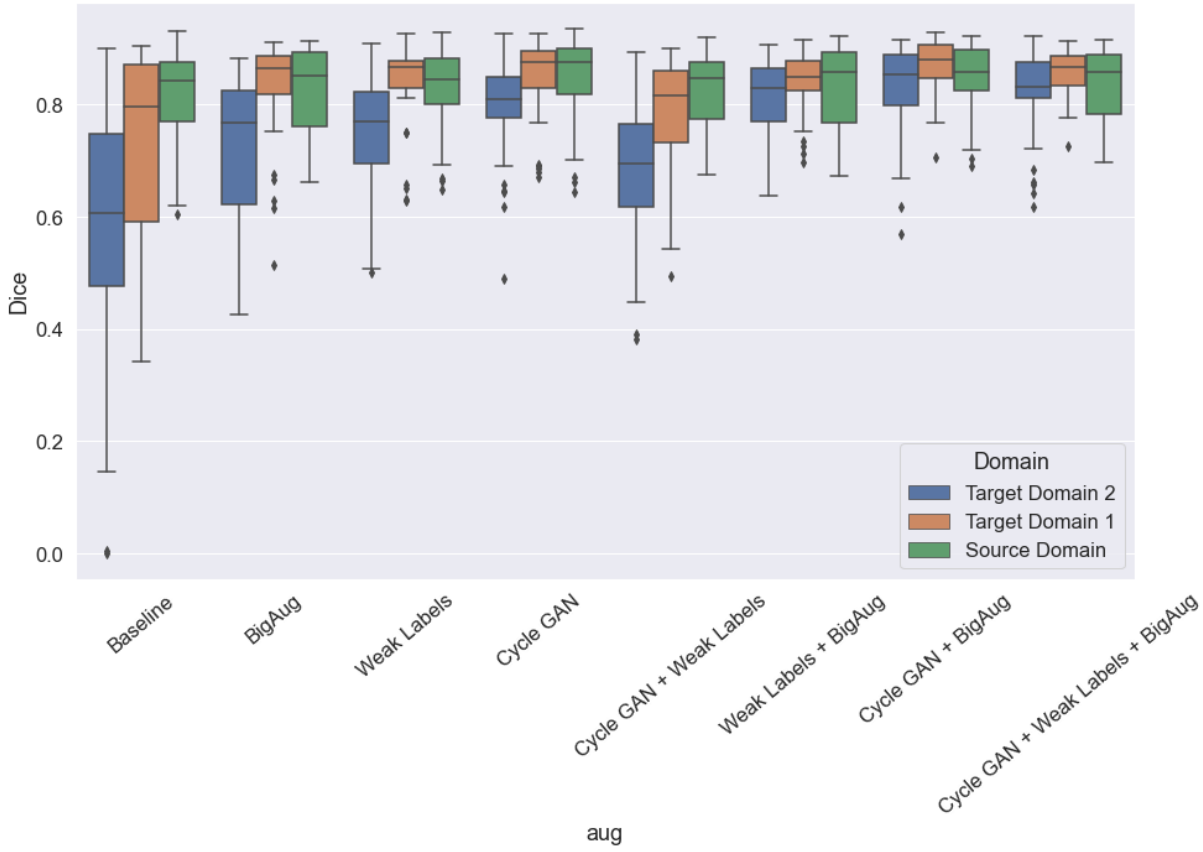


(a) DICE scores

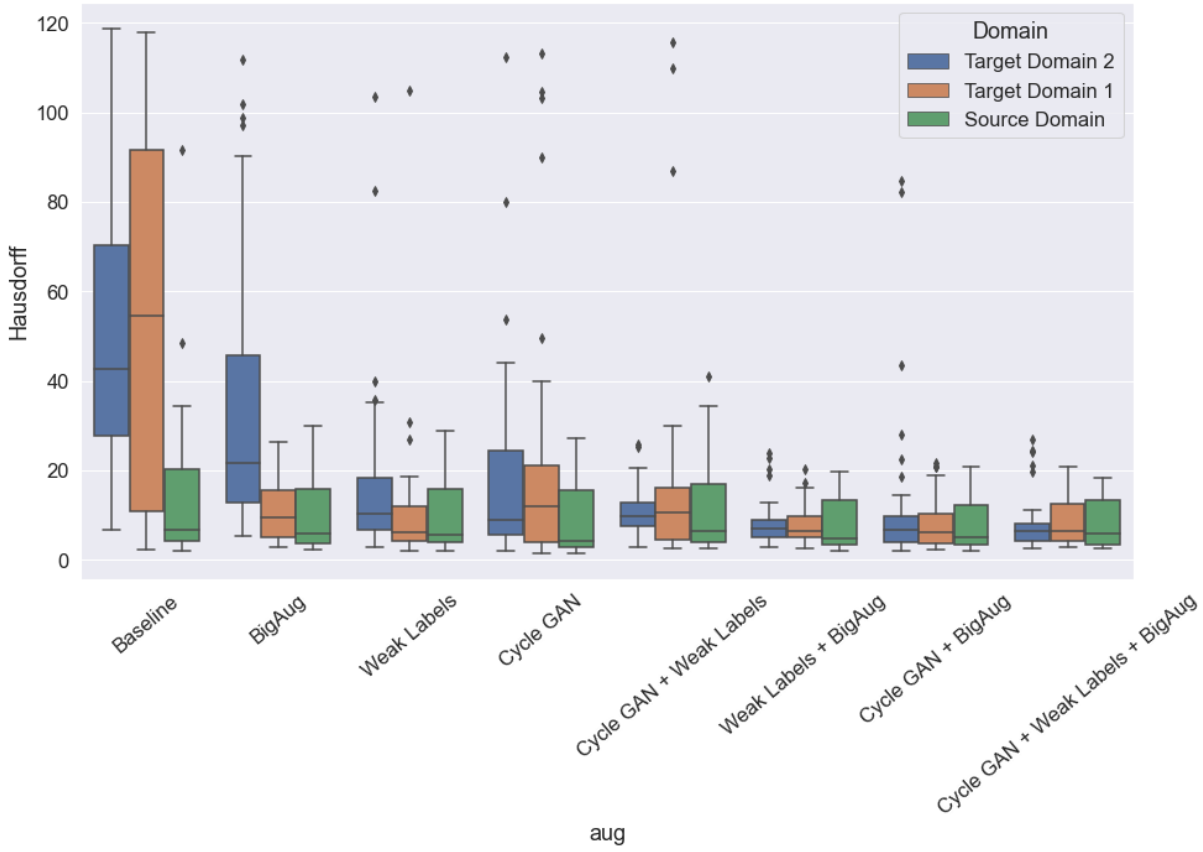


(b) $H_{95\%}$ distances

Figure 20: Score comparison for the urinary bladder on the test sets



(a) DICE scores



(b) $H_{95\%}$ distances

Figure 21: Score comparison for the rectum on the test sets

5.5 Qualitative evaluation of augmentation methods

While the numerical metrics provide some insight to model performance, they do not give the full picture. Therefore, the predicted segmentation masks of the models were studied manually in order to get a better understanding of the results. This section provides a summary of the key observations from this study, and shows a number of samples of predicted segmentations. As the report is limited to 2 dimensions only slices of these samples are shown, and the reader should keep in mind that this is not an ideal way of presenting 3D data.

In general, the quality of the segmentations are correlated to the metrics discussed above. This means that for the best models, the frequency of both types of errors discussed in Section 5.1 is significantly reduced. Figure 22 shows the same slices as Figure 12, but with the predicted segmentation of the model that performed best on each respective domain, on average. That is, for $\mathcal{D}_S$ and $\mathcal{D}_{T1}$ the predictions are from the 'Cycle GAN + BigAug' model, while the predictions on $\mathcal{D}_{T2}$ are from the model where all three methods are combined. Comparing with the baseline results (Figure 12) the improvements are clear, expecially for the two target domains. It is also clear that the predicted segmentations are in no way perfect.

Besides the general observation above, an interesting property of the models trained on noisily labeled samples is highlighted by the manual inspection; the tendency of these models to predict several independent structures is significantly reduced compared with the models that were not trained using the weak labels. Additionally, the boundaries of the RoI borders are in general smoother, and the RoI shapes appear more realistic (for example, predicted RoI masks do not have holes in them). This RoI-shape regularization mainly occurs in the target domains, especially $\mathcal{D}_{T2}$, and is visualized for one $\mathcal{D}_{T2}$ patient in Figure 23. For $\mathcal{D}_S$, the RoI shapes predicted in the baseline experiments are already quite regular, and no clear difference is seen between these results and the results for training on weak labels.

However, for a training setup that already performs well on a certain patient, adding weakly labeled samples to the training seems to not have a positive effect as it increases the distances between predicted RoI boundaries and their true boundaries. This is illustrated in the bottom row of Figure 23.

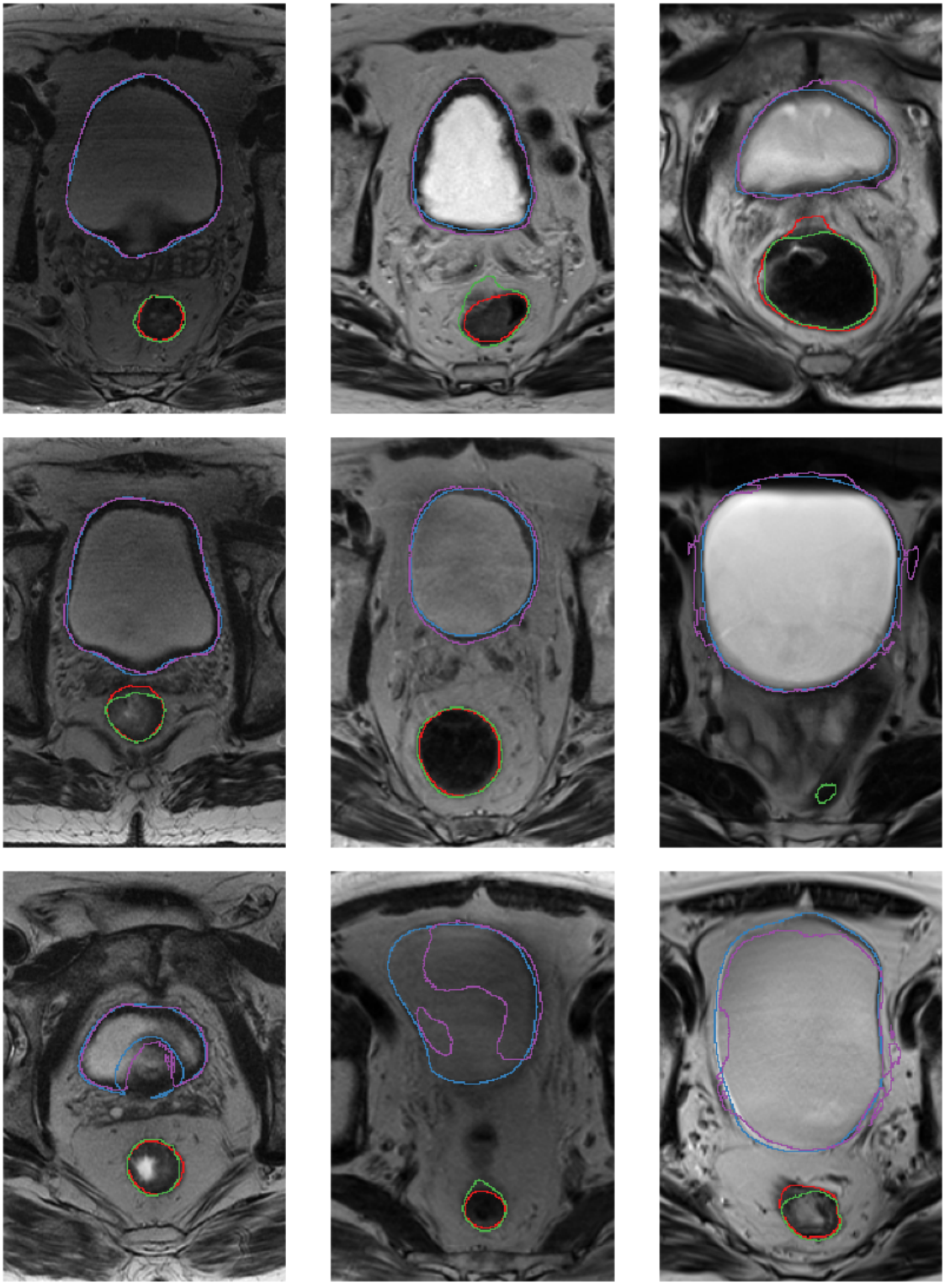
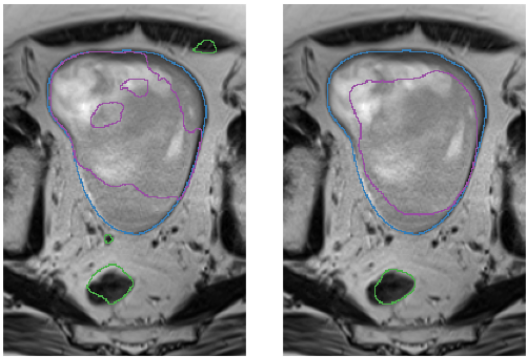
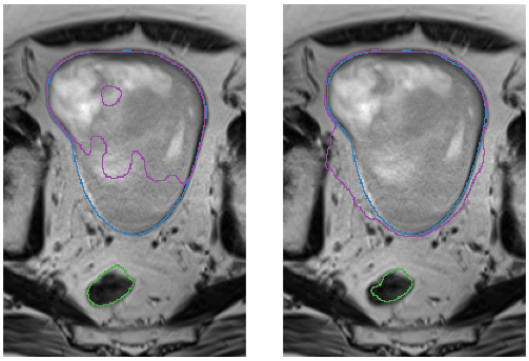


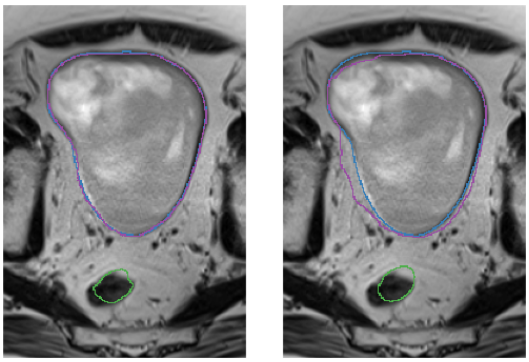
Figure 22: Sample of the segmentation results for the best models on each domain, with columns corresponding to $\mathcal{D}_S$ (left, 'Cycle GAN + BigAug' model), $\mathcal{D}_{T1}$ (middle, 'Cycle GAN + BigAug model') and $\mathcal{D}_{T2}$ (right, 'All' model). Ground truth: rectum = red, bladder = blue. Predictions: rectum = green, bladder = purple. Comparing to the baseline results (see Figure 12) it is clear that segmentation performance is greatly improved, especially on $\mathcal{D}_{T1}$ and $\mathcal{D}_{T2}$.



(a) Baseline (left) and Weak Labels (right) segmentations



(b) BigAug (left) and BigAug + Weak Labels (right) segmentations

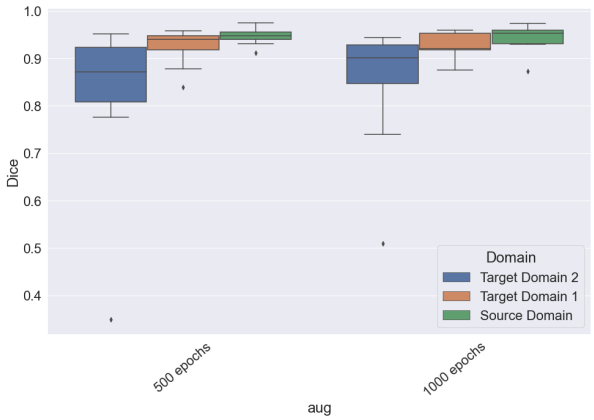


(c) Cycle GAN + BigAug (left) and Cycle GAN + BigAug + Weak Labels (right) segmentations

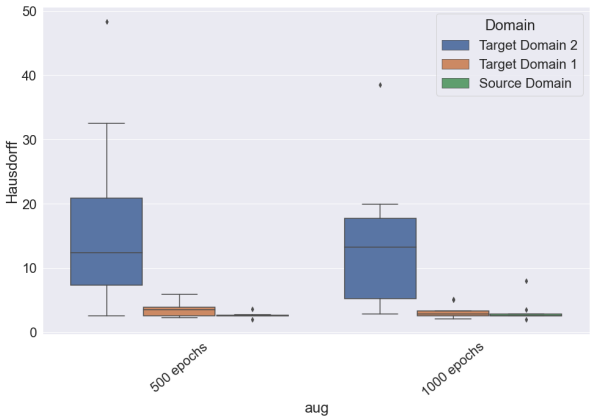
Figure 23: Different segmentations of the same $\mathcal{D}_{T2}$ slice, illustrating the regularizing properties of using weak labels for training of the segmentation model. For each row, the result of a model trained without weak labels is shown to the left. To the right, the result of the same training setup is shown, with the difference that noisily labeled $\mathcal{D}_{T2}$ images were included. In the top two rows, it is clear that the model trained on weak labels predict more realistic bladder shapes, and in general only one structure is predicted per RoI. In the bottom row, however, it can be seen that the ability to follow the true RoI contours is decreased when predictions are already good.

5.6 On the training time of combined approaches

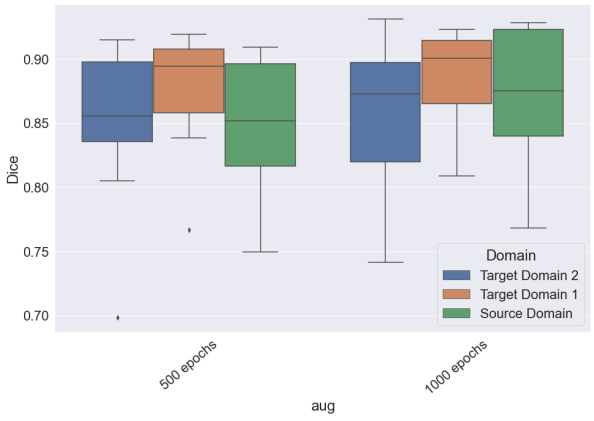
The number of iterations to be used for the training of the segmentation model was determined largely from the validation results on the baseline experiments, as these were the first ones to be carried out. As the augmentation pipeline got more complex. It was noted that the time before convergence on the $\mathcal{D}_S$ validation set was prolonged, and for some of the experiments it is not sure whether the validation loss had fully plateaued, suggesting that better results would be obtained if the model was trained for more iterations. However, this was not carried through due to time constraints. This does not effect the conclusions of the thesis as the main questions are related to whether the augmentation techniques reduce the domain shift error, which they clearly do. What remains unclear is *how* much better the models would be if they were trained for a longer period of time. To get a hint of final model performance, a training was carried out using the 'Cycle GAN + BigAug' setup for twice the number of epochs, the results of which are shown in Figure 24. From this, we draw the conclusion that model performance will likely be improved by longer training times.



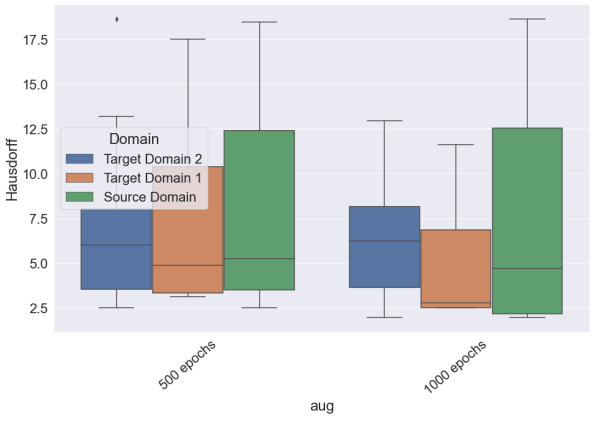
(a) Urinary bladder Dice scores



(b) Urinary bladder $H_{95\%}$ distances



(c) Rectum Dice scores



(d) Rectum $H_{95\%}$ distances

Figure 24: Results of training a segmentation model with the BigAug pipeline and Cycle GAN generated data, after 500 and 1000 epochs. In general, metrics keep improving after the 500:th epoch for all domains

6 Discussion

6.1 Effects of the augmentation methods on segmentation performance

Firstly, it is clear that all three methods examined lead to increasing ability of the segmentation model to generalize to unseen domains, as illustrated by Tables 4-6 and Figures 20-21. Adding to this, there seems to be an even larger gain in performance to be achieved by combining the methods instead of using only one of them. An observation to be made here is that neither the Cycle GAN images nor the weak labels are perfect - that is, the weak labels are often misaligned with the true RoI borders, and the cycle GAN sometimes introduces artefacts such as new structures in the synthetic data. This means that one should be careful about using the weak labels or Cycle GAN generated data on their own in a real world clinical setting, for example using the weak labels for treatment planning, or the synthetic data for automated diagnostics. Nonetheless, the experiments show that both weak labels and Cycle GAN-generated data are valuable in a data augmentation setting, where the mistakes made are of less importance as the data is only used for model training.

For the BigAug pipeline, performance on both target domains increase while performance on $\mathcal{D}_S$ seems largely unaffected. It is somewhat surprising that the augmentation pipeline does not improve performance on the source domain, as such augmentation steps have been known to significantly increase model generalization ability on other computer vision tasks.

Including $\mathcal{D}_{T2}^{synthetic}$ in the training sets improves segmentation performance on all domains. For $\mathcal{D}_S$, only doing this augmentation step yields the best result on all metrics but the rectum $H_{95\%}$ distance, which is also the only metric which was improved by the BigAug pipeline on this domain. For $\mathcal{D}_{T2}$, including $\mathcal{D}_{T2}^{synthetic}$ yields better results than including weak labels or using the BigAug pipeline, while for $\mathcal{D}_{T1}$ the 'Cycle GAN' results are comparable to the 'BigAug' results. This can be taken as an indication that the Cycle GAN approach is more domain-specific than the BigAug approach, i.e. the performance gains will likely be higher on the domain that the Cycle GAN was trained on than on other domains. However, the improvements on $\mathcal{D}_S$ speak against this interpretation of the results.

Disregarding the fine-grained details, it is still clear that both the BigAug and the Cycle GAN approaches provide improvements in the segmentation results on the target domains while not affecting the source domain negatively, and that combining the two methods yield even better results. With the weak labels approach, however, the results are more ambiguous. In favor of the approach, we note that including noisily labeled $\mathcal{D}_{T2}$ data in the training improves the result compared to the baseline on both target domains, and combining this with the BigAug pipeline yields better results than only using BigAug augmentation. However, for $\mathcal{D}_S$ and $\mathcal{D}_{T1}$ the 'Cycle GAN + BigAug' experiments in general yield better results than also including the weakly labeled $\mathcal{D}_{T2}$ data. On the other hand, for $\mathcal{D}_{T2}$ the best results on the DICE and $H_{95\%}$ metrics are obtained by using all three methods. A possible explanation to these observations might

be found by inspecting the predicted segmentations qualitatively, as described in section 5.5. As shown in Figure 23, the inclusion of weakly labeled samples in the training set acts as an RoI-shape regularizer, i.e. the number of independent structures predicted for each RoI is reduced, and the predicted shapes are less sprawling. This effect greatly reduces the number of predictions that are catastrophically bad. However, as seen in the bottom row of Figure 23, the performance gain achieved by including weakly labeled samples for training diminishes when model performance improves, as the inclusions of these samples can lead to reduced ability of the model to correctly predict RoI boundaries. This is not a surprising result considering the nature of the weak labels, which are good at locating the RoI:s but do not follow boundaries well.

Thus, the answer to the question as to whether weakly labeled samples should be added to the 'Cycle GAN + BigAug' method or not depends on the purpose of the segmentation model. If the purpose is to reduce the number of catastrophic predictions, but the demands on segmentation quality are not too high, including weak labels is likely a beneficial augmentation step. However, if the predicted segmentations need to be on par with the performance of clinical practitioners, then excluding weak labels from the training is likely to increase the chances of adequate segmentations, at the cost of bad predictions becoming even worse.

6.2 What causes the domain shift error?

Comparing between $\mathcal{D}_{T1}$ and $\mathcal{D}_{T2}$, there are strong indications that $\mathcal{D}_{T1}$ is closer to the source domain. This can be seen in the data set statistics of Figures 4, 5 and 6, as well as by the fact that the segmentation results on $\mathcal{D}_{T1}$ are better than on $\mathcal{D}_{T2}$ in nearly all experiments. The results of the augmentation experiments all indicate that the domain-shift error is can be reduced more for $\mathcal{D}_{T1}$, suggesting that the statistics used in Figures 4-6 can be indicative of the segmentation model performance.

However, the question as to what properties of the MR images are the most contributing factors to the observed domain shift error remains largely unanswered. In order to answer this question, future studies might try to find MRI properties that correlate with segmentation model performance. For this thesis, the test set was deemed to small to get reliable answers to this question, and manual inspection of the images did not result in the identification of any clear patterns.

7 Conclusions and future work

Concluding the discussion above in relation to the questions that this thesis aimed to answer, we note

- All three augmentation methods - that is, classical data augmentation (BigAug), synthetic data generation (Cycle GAN) and weak labels - provide improvements on the baseline segmentations of the unseen domains.
- The best results are achieved by combining several of methods, and which combination is best depends on the purpose of the segmentation model. If the most important goal is to reduce the number of catastrophic predictions then all of the three methods should be combined. If, however, the requirements on segmentation quality is high, then excluding weakly labeled samples from training is likely to increase the number of acceptable segmentations.
- For the best performing models, the domain difference in the studied metrics for delineation of the rectum is nearly undetectable, while the predicted bladder contours are still better on the seen (source) domain.
- A number of image features (e.g. relative RoI intensities) have been identified where there are high inter-domain variations, and which seem to be correlated to segmentation performance. However, it has not been determined which of these actually cause to the domain shift error, and there are likely a large number of such features that remain to be studied. Therefore, there is no answer to what domain differences affect the domain shift error most.

This thesis has studied the problem of inter-domain performance variations of deep learning segmentation models using data from three separate domains. An interesting area for future research would be to extend the number of domains studied to further understand how the augmentation methods discussed generalize. As the test set sizes in this thesis are very small, the result quality of such research would likely be improved if the data sizes within each domain were larger. This would increase both robustness of the results and overall model performance. Furthermore, it would be of interest to explore different set-ups for training the Cycle GAN, for example reweighting the loss functions and including data augmentation to avoid the overfitting behavior of the discriminators.

7.1 Final remarks on clinical relevance

While the methods studied result in significantly improved segmentation of unseen data domains, the results do not reach the same level as they do on the source domain for all RoI:s. Furthermore, comparing with the human level performance in Table 5, it is clear that the models perform worse than clinical experts on all domains. As the results lag behind human performance even on the source domain, it is likely that more annotated data is needed to successfully train deep learning segmentation models that reach the performance level of clinical experts. However, the results presented indicate that it might

not be necessary to acquire labeled data from all domains that one needs the model to perform well on. For the domains studied here, it is reasonable to believe that labeled $\mathcal{D}_{T2}$ data would be enough to improve performance sufficiently on both target domains. An interesting question for future research is then to find a way of determining from which domains labeled data would improve general segmentation performance the most, as this could provide a way to allocate labelling resources efficiently. As the methods discussed improved general performance on all domains, it is likely that they would be useful for training segmentation models on larger data sets as well.

References

[1] Amy Zhao, Guha Balakrishnan, Frédo Durand, John V. Guttag, and Adrian V. Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. *arXiv:1902.09383 [cs]*, April 2019. arXiv: 1902.09383.

[2] Magnetic Resonance Imaging. In Steven D. Waldman and Robert S. D. Campbell, editors, *Imaging of Pain*, pages 19–21. W.B. Saunders, Philadelphia, January 2011.

[3] Amir M. Owrangi, Peter B. Greer, and Carri K. Glide-Hurst. MRI-only treatment planning: benefits and challenges. *Physics in medicine and biology*, 63(5):05TR01, February 2018.

[4] Emilia Persson, Christian Jamtheim Gustafsson, Petra Ambolt, Silke Engelholm, Sofie Ceberg, Sven Bäck, Lars E. Olsson, and Adalsteinn Gunnlaugsson. MR-PROTECT: Clinical feasibility of a prostate MRI-only radiotherapy treatment workflow and investigation of acceptance criteria. *Radiation Oncology (London, England)*, 15(1):77, April 2020.

[5] Neeraj Sharma and Lalit M. Aggarwal. Automated medical image segmentation techniques. *Journal of Medical Physics / Association of Medical Physicists of India*, 35(1):3–14, 2010.

[6] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *Journal of Digital Imaging*, 32(4):582–596, August 2019.

[7] Hao Guan and Mingxia Liu. Domain Adaptation for Medical Image Analysis: A Survey. *arXiv:2102.09508 [cs, eess]*, February 2021. arXiv: 2102.09508.

[8] Donald W. McRobbie. *MRI from Picture to Proton*. Cambridge University Press, Cambridge ; New York, 3rd edition edition, August 2017.

[9] Kenneth S. Krane. *Introductory Nuclear Physics*. New York, November 1987.

[10] Christopher J. Foot. *Atomic Physics*. Oxford University Press, Oxford ; New York, 1st edition edition, February 2005.

[11] Tufve Nyholm, Stina Svensson, Sebastian Andersson, Joakim Jonsson, Maja Sohlin, Christian Gustafsson, Elisabeth Kjellén, Karin Söderström, Per Albertsson, Lennart Blomqvist, Björn Zackrisson, Lars E. Olsson, and Adalsteinn Gunnlaugsson. MR and CT data with multiobserver delineations of organs in the pelvic area-Part of the Gold Atlas project. *Medical Physics*, 45(3):1295–1300, March 2018.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597 [cs]*, May 2015. arXiv: 1505.04597.

[13] https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/Adam | TensorFlow Core v2.4.1.

[14] Connor Shorten and Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60, July 2019.

[15] Ling Zhang, Daguang Xu, Ziyue Xu, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford Wood, Holger Roth, and Andriy Myronenko. Generalizing Deep Learning for Medical Image Segmentation to Unseen Domains via Deep Stacked Transformation. *IEEE Transactions on Medical Imaging*, PP:1–1, February 2020.

[16] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, January 2018.

[17] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep Learning is Robust to Massive Label Noise. *arXiv:1705.10694 [cs]*, February 2018. arXiv: 1705.10694.

[18] Juan Eugenio Iglesias and Mert R. Sabuncu. Multi-Atlas Segmentation of Biomedical Images: A Survey. *Medical image analysis*, 24(1):205–219, August 2015.

[19] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*, June 2014. arXiv: 1406.2661.

[20] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, Venice, October 2017. IEEE.

[21] Casey Chu, Andrey Zhmoginov, and Mark Sandler. CycleGAN, a Master of Steganography. *arXiv:1712.02950 [cs, stat]*, December 2017. arXiv: 1712.02950.

[22] 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. arXiv: 1809.05231.

Appendix A: Learned transformations to generate synthetic data

Initially, another method for creating synthetic training data was used than Cycle GAN was tried out. However, this method proved difficult to get to work properly on the anatomical region studied in the thesis. Included below is a description of the method as well as a visualization of the initial results.

A.1 Method description

A potential drawback with the Cycle-GAN, is that it is not able to synthesize new geometries in the images, while still keeping accurate to the labels. However, Zhao et. al. [1] propose another deep learning-based augmentation technique that might remedy this. The technique is based on learning two transformations that are applied to the images in $\mathcal{D}_S^{train}$.

The first transformation τ_s is a spatial deformation of the image which is represented by a voxel-wise displacement field

$$\mathbf{u} : \mathbb{R}^{w \times d \times h} \longrightarrow \mathbb{R}^{w \times d \times h \times 3} \tag{12}$$

i.e. a displacement field in three dimensions for each voxel in the image. This vector field defines the deformation function

$$\phi = \mathrm{id} + \mathbf{u} \tag{13}$$

where id is the identity function. The function $\mathbf{u}$ is parameterized by a 3D U-Net, which is trained to maximize the similarity between the source and target images (as measured by a local cross-correlation function) while ensuring that $\mathbf{u}$ is smooth by adding a regularization term consisting of the norm of the spatial gradients of $\mathbf{u}$, forming

$$\mathcal{L}_{\tau_s} = \mathcal{L}_{sim} + \lambda \mathcal{L}_{smooth} \tag{14}$$

where λ is a weighting parameter [22].

The second transformation is called an *appearance transformation* and consists of a voxel-wise addition:

$$\tau_a(\mathbf{x}) = \mathbf{x} + \boldsymbol{\psi} \tag{15}$$

ψ is also parameterized with a U-Net, which is trained to optimize the voxel-wise squared difference in intensities between the source and target image, while discouraging drastically different intensity changes within a specific ROI within the target image [1].

As for the Cycle-GAN experiments, the transformation U-Nets are trained on $\mathcal{D}_S^{train}$ and $\mathcal{D}_T^{unlabeled}$. The resulting transformations are then applied on $\mathcal{D}_S^{train}$ to create synthetic training data. Contrary to the Cycle-GAN, the labels of $\mathcal{D}_S^{train}$ are also transformed using the spatial transformation, meaning that the synthetic data set is

$$\mathcal{D}_{T2}^{synthetic} = \left\{ \left(\tau_s \left(\tau_a(\mathbf{x}_i^S) \right), \tau_s(\mathbf{y}_i^S) \right) \right\}_{i=1}^{n_S^{train}} \tag{16}$$

The segmentation network is trained on the combined data set $\mathcal{D}_S^{train} \cup \mathcal{D}_{T2}^{synthetic}$.

A.2 Learning the spatial transformation

After some experimentation, the learned transformation approach was abandoned as it proved too difficult to train a U-Net to generate vector fields that gave acceptable result when applied to source image data. Figure 25 shows a sample of the training results, which make clear that the spatial deformation is not succesful in warping the source image to imitate the target.

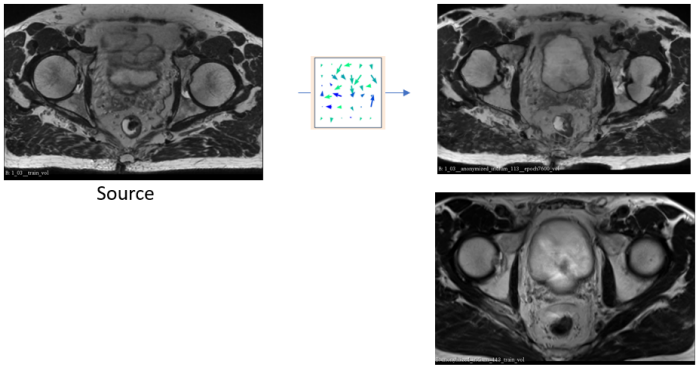
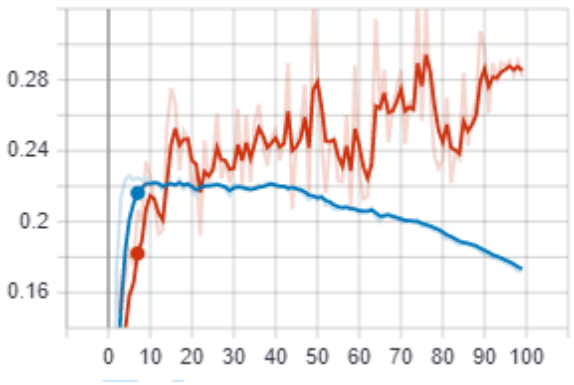
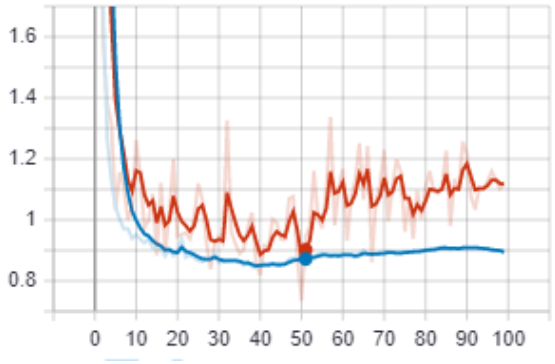


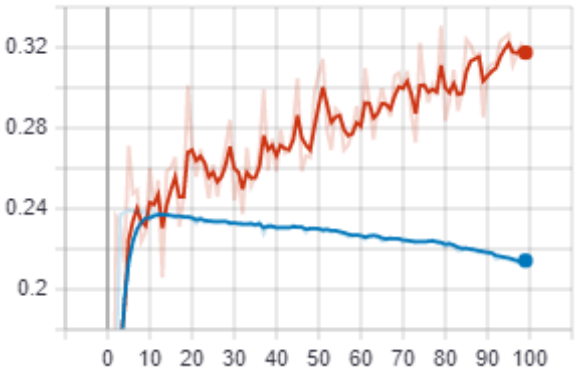
Figure 25: A sample of a $\mathcal{D}_{T1}$ (Source) image, and the result of applying a learned transformation to it, to imitate the geometry of the target image from $\mathcal{D}_{T2}$. The applied vector field does not do a good enough job at imitating the target geometry, and this method to generate synthetic data was never used to train a segmentation network.



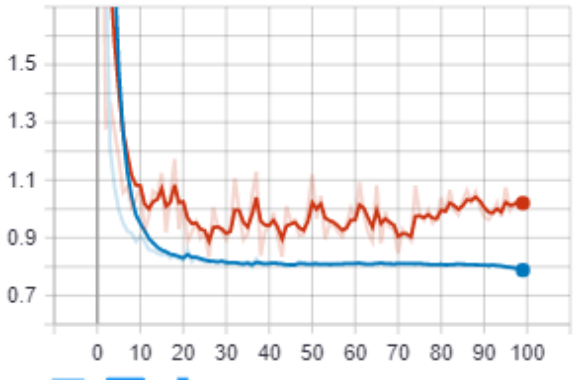
(a) $\mathcal{D}_S$ discriminator loss



(b) $\mathcal{D}_S$ generator loss



(c) $\mathcal{D}_{T2}$ discriminator loss



(d) $\mathcal{D}_{T2}$ generator loss

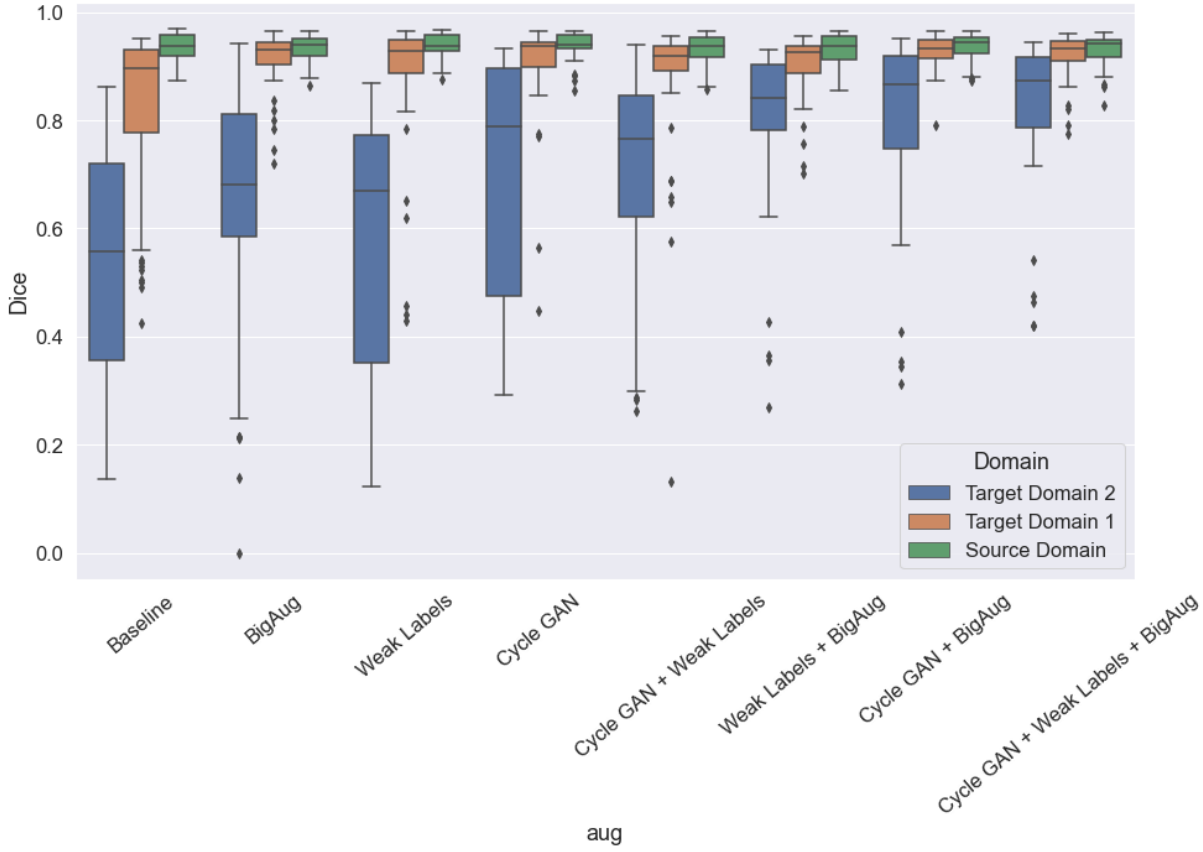
Figure 26: Training (blue) and validation(red) loss curves for the training of Cycle GAN, x-axes represents number of epochs. Clearly, the discriminators overfit to the training data

Appendix B: Cycle GAN training curves

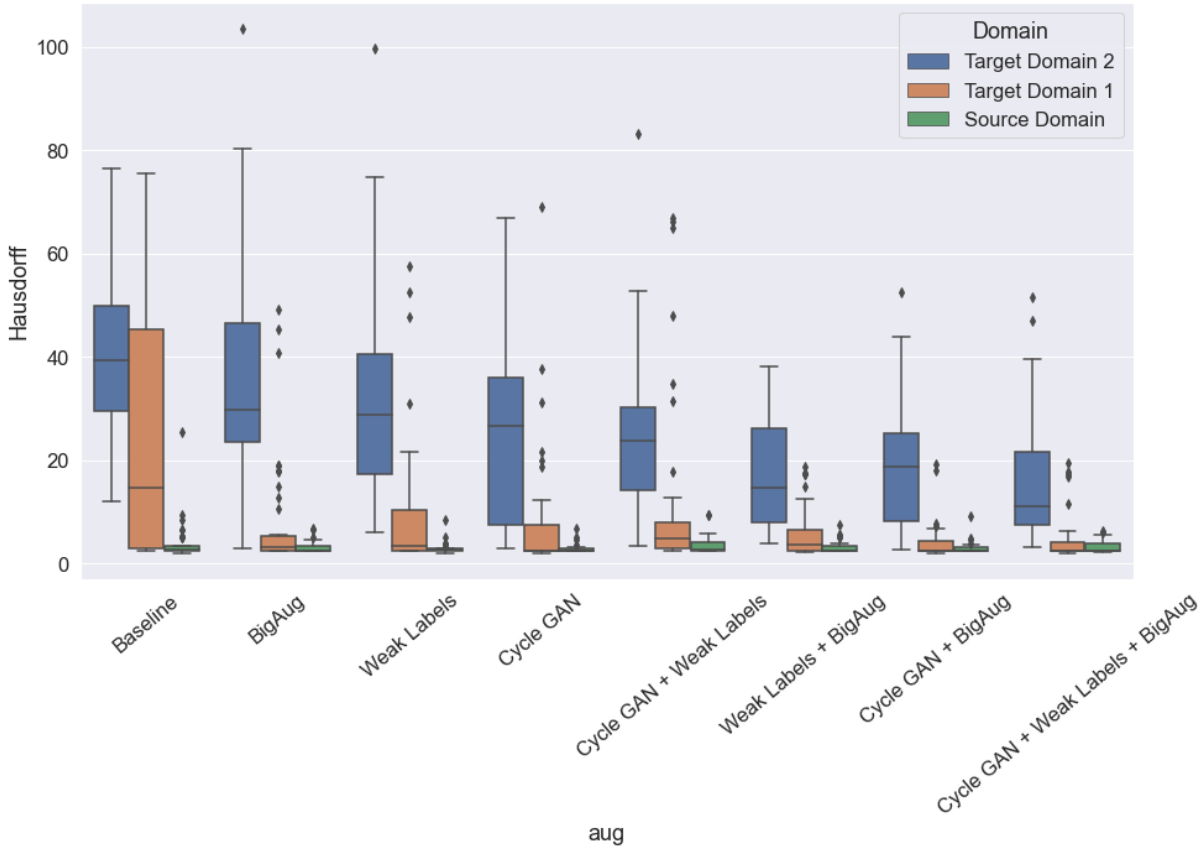
Figure 26 shows the loss functions for the training of the Cycle GAN used to generate $\mathcal{D}_{T2}^{synthetic}$. As discussed in Section 5.3, both discriminators overfit to the training data quite drastically. When the discriminators are overfit, it is reasonable to assume that the generators do not learn to generate general target domain data, but rather try to imitate the training set of the target domain. It would be interesting to investigate methods to reduce this overfitting behavior, however training a segmentation network on synthetic data generated by the overfitted Cycle GAN still improves segmentation performance significantly as seen in Section 5.4.

Appendix C: Experimental results on validation sets

Figures 27 and 28 show the score distributions on the validation sets. As stated, optimization on the validation sets of the target domains was not performed extensively, meaning that the validation results should be fairly indicative of model performance on unseen test data. For the source domain, the model choice was based on validation loss, i.e. it is likely that the models are overfitted to the validation set. In general, the patterns in these results are the same as for the test sets. However, the performance on $\mathcal{D}_{T2}$ is generally lower on this validation set, and somewhat higher for $\mathcal{D}_{T1}$

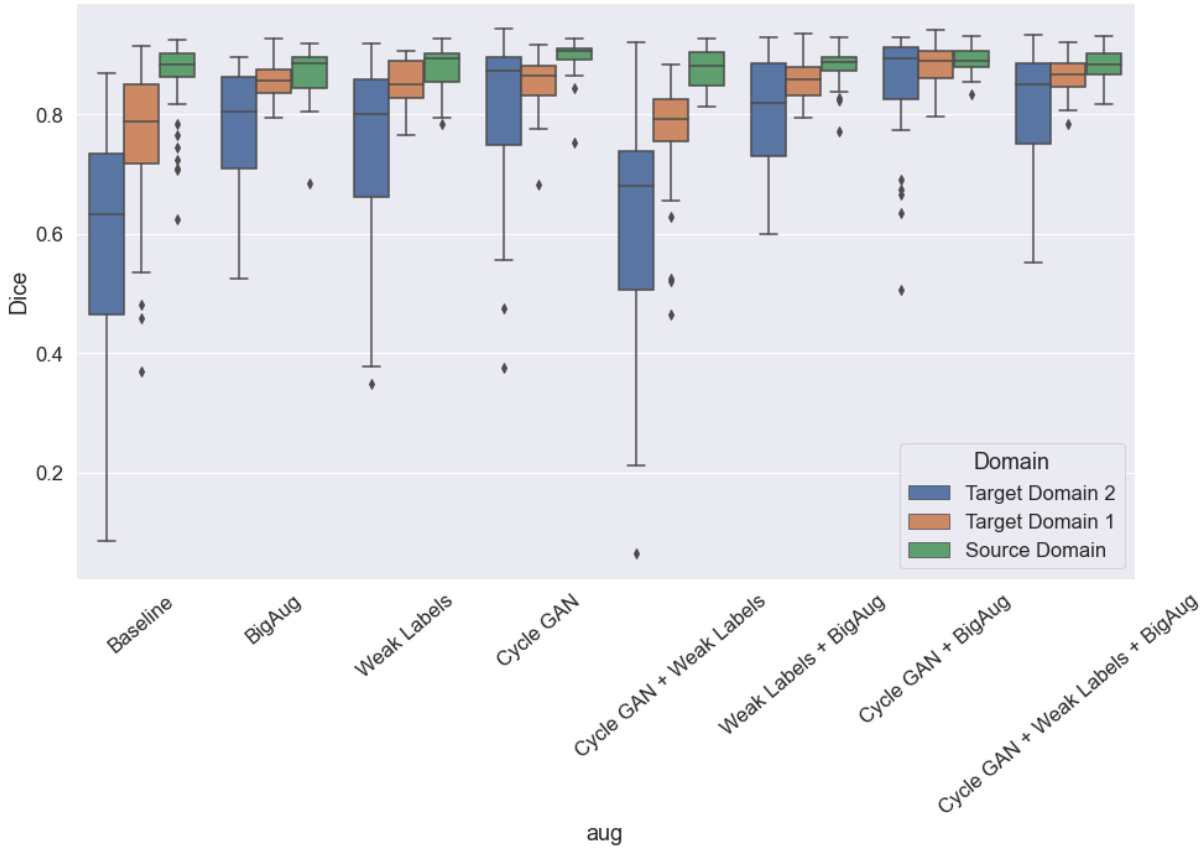


(a) DICE scores

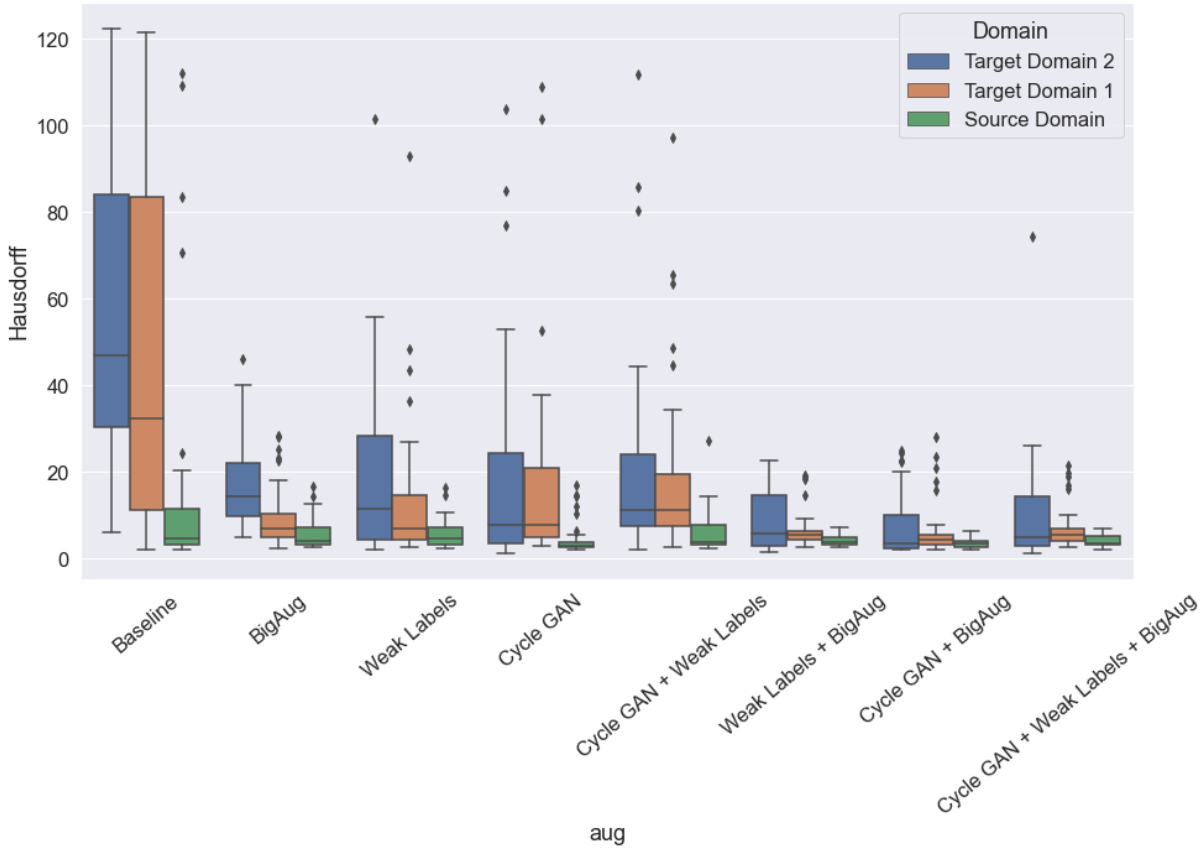


(b) Hausdorff 95 scores

Figure 27: Score comparison for the urinary bladder on the validation sets



(a) DICE scores



(b) Hausdorff 95 scores

Figure 28: Score comparison for the rectum on the validation sets

Master's Theses in Mathematical Sciences 2021:E26
ISSN 1404-6342

LUTFMA-3445-2021

Mathematics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/