Universiteit Leiden

Master Computer Science

Semi-Supervised Consistency Training for Image Segmentation in 3D CT Scans

Name:	Yichao Li
Student ID:	s2372029
Date:	March 29, 2021
Specialisation:	Data Science
1st supervisor:	Michael Lew
2nd supervisor:	Marius Staring
3rd supervisor:	Mohamed Elmahdy

Master's Thesis in Computer Science

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

Acknowledgement

The HMC dataset with contours was collected at Haukeland University Hospital, Bergen, Norway, and was provided to us by responsible oncologist Svein Inge Helle and physicist Liv Bolstad Hysing. The EMC dataset with contours was collected at Erasmus University Medical Center, Rotterdam, The Netherlands, and was provided to us by radiation therapist Luca Incrocci and physicist Mischa Hoogeman. We are grateful for their contributions. Elastic deformation was implemented with code from Huang et al. [1] and image samplers from NiftyNet [2] were used, for these we are also very grateful.

I would like to thank Prof. Marius Staring and Mohamed Elmahdy for their ideas, guidance and the amount of time they have contributed to supporting this project from a medical application perspective, and Prof. Michael Lew from a more general computer science perspective. I would also like to thank Margherita Grespan and Iris Cornelissen for their company during the many months it took to complete this work.

Contents

Abstract

Deep supervised models often require a large amount of labelled data, which is difficult to obtain in the medical domain. Therefore, semi-supervised learning (SSL) has been an active area of research due to its promise to minimize training costs by leveraging unlabelled data. Previous research have shown that SSL is especially effective in low labelled data regimes, we show that outperformance can be extended to high data regimes by applying Stochastic Weight Averaging (SWA), which incurs zero additional training cost. Furthermore, we also conclude that larger-than-realistic transformations are the most beneficial. Our model was trained on a prostate CT dataset and achieved improvements of 0.12 mm, 0.14 mm, 0.32 mm, and 0.14 mm for the prostate, seminal vesicles, rectum, and bladder respectively, in terms of median test set mean surface distance (MSD) compared to the supervised baseline in our high data regime.

Keywords— semi-supervised learning, Image Segmentation, Consistency Loss, Stochastic Weight Averaging (SWA), Convolutional Neural Networks (CNN), Adaptive Radiotherapy

1 Introduction

Supervised deep learning models have proven to be effective in many computer vision problems, but training such models for practical applications requires a large amount of labelled data [3]. These are especially difficult to obtain in the medical domain due to the reliance on highly specialised personnel and patient confidentiality considerations. Therefore, semi-supervised learning (SSL), which use both labelled and unlabelled data, are especially relevant and a number of methods for medical image segmentation have already been proposed [4, 5, 6, 7, 8].

The idea of transformation-consistent SSL, as proposed by Sajjadi *et al.* [9] for classification, is that one can apply transformations or perturbations to the input image without changing the label. For unlabelled images, this means that the network should predict the same label before and after the transformation. This idea has already been adapted for medical image segmentation [4, 5], so the main contribution of this paper is to show that Stochastic Weight Averaging (SWA) [10], which averages a sample of network weights along the convergence path, further improves performance without incurring additional training cost. This is especially significant in high labelled data regimes, in which previous research have shown that SSL does not outperform substantially [4, 5]. Moreover, we also examine the effects of different transformation procedures.

The clinical context of this paper is the treatment of prostate cancer with radiotherapy, where we maximize the treatment dose to the target organs (prostate and seminal vesicles), while minimizing the dosage to the surrounding organs-at-risk (OARs) (bladder and rectum). Therefore, it is crucial to precisely segment target organs and OARs in order to avoid treatment related complications [11].

Note on Notation Capital letters in bold denote tensors of dimension greater than 1, e.g. $\mathbf{X}$ is the input tensor, while italic capital letters indicate sets, e.g. X_u is the set of unlabelled input images.

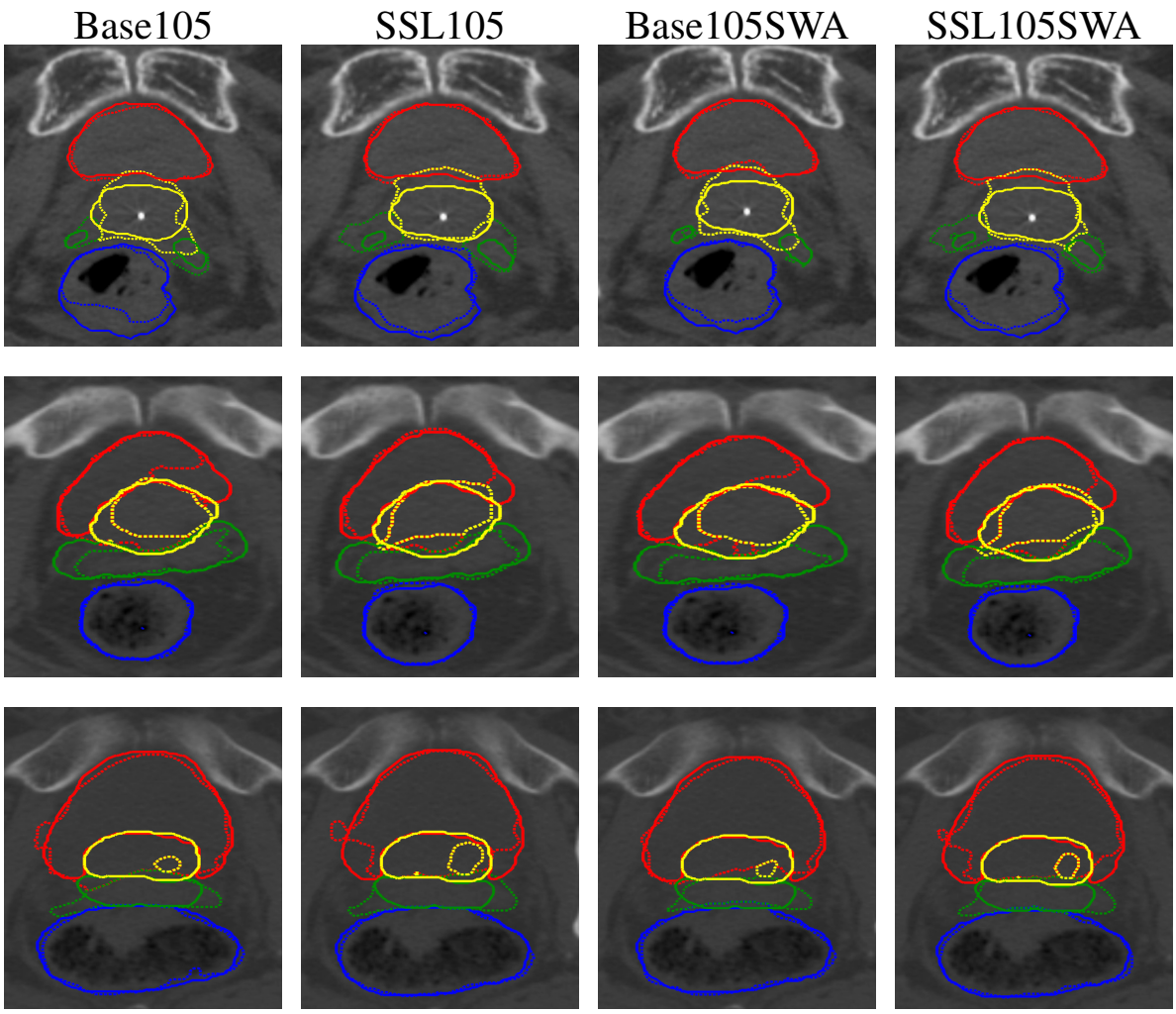


Figure 1: Example segmentations from the high data regime for different networks. From top to bottom, the images are selected from the first, second and third quartiles in terms of the prostate MSD of the SSL105SWA network. The solid lines are groundtruth delineations by a radiation oncologist and the dotted lines are produced by the networks. Red, yellow, blue, and green represent the bladder, prostate, rectum, and seminal vesicles, respectively.

2 Related Work

In existing work on classification [9], the idea of consistency also extends to perturbing the intermediate representations in the network, by using dropout [12] and randomised max-pooling [13], but this paper focuses on transformation of the input only. The Π-model introduced in [14] is based on the same principle, but expands the idea with Temporal Ensembling, which takes an exponential average of the past predictions of the current image in the previous epochs as unlabelled target, thus taking advantage of more stable predictions. Mean Teacher [15] further builds on this idea by averaging the weights instead, which allows unsupervised targets to update after each iteration, as opposed to every epoch in the case of Temporal Ensembling. Xie *et al.* [16] contributes to the discussion by examining the effect of different input transformations on performance, which is the motivation behind the experiment of Section 4.3.

It is also worth noting that consistency is just one of many approaches to semi-supervised learning. One early branch of inquiry used unlabelled data to model a joint distribution of the input data and labels, i.e. a generative model [17]; the disadvantage with this approach is that discriminative models performs better as the amount of training data increases [18]. A very simplistic idea is to use the network's own predictions on the unlabelled set as the target [19], which encourages more confident predictions; the intuition is that this reduces the entropy of the prediction, which is consistent with the decision boundary being in the low density regions of the input space. Further work incorporates entropy minimisation more explicitly into the model specification [20]. Early works with convolutional neural networks focused on using unlabelled data to pretrain the model [21, 22], thus starting with more promising initial weights to speed up convergence; however, other works [23, 24] have shown that random initialisation can still produce state of the art results. A more recent idea that has been shown to work well in classification settings is Mixup [25], which encourages smoother class boundaries [26], hence reduces overfitting.

Cheplygina *et al.* [27] provided a survey of earlier research on SSL for medical image segmentation and classified them into two approaches. The idea of closeness assumes that samples close to each other in the sample space might have the same label and it is embodied by self-training, which propagate high-confidence labels to unlabelled images as training progresses. The second approach is the idea of clustering, which assumes that clusters of samples might have the same label; this is typically expressed by graph- and SVM-based methods that try to place class boundaries in low density regions.

The idea of transformation-consistency has been widely explored by recent research. In addition to Bortsova *et al.* and Li *et al.* [4, 5], on which our method is based, the Mean Teacher model [15] from classification has also been successfully adapted for image segmentation [6, 8]. Liu *et al.* [7] extended the idea of consistency to include group similarity measures, instead of focusing only on the individual-sample level. Compared to earlier research, these transformation-consistent methods have a lot in common with self-training [28]. Finally, Athiwaratkun *et al.* [29] demonstrated that SWA is especially beneficial for consistency-based SSL in a classification setting, hence it is potentially applicable to all similar methods in a segmentation setting, even though our methodology is the most similar to [4, 5].

3 Methodology

3.1 Context and Motivations Behind Design Choices

The semi-supervised method in this paper adapts the image classification framework in [9, 14] for segmentation, but the intermediate representations were not perturbed, because the existing supervised baseline does not use any stochastic perturbation layers (although batch normalisation is used, this is intended for data standardisation rather than perturbation). Secondly, instead of a λ-schedule, Xie *et al.* [16] suggests a more sophisticated procedure, whereby a confidence threshold is defined for unlabelled predictions and the unsupervised loss is only calculated on individual inputs whose softmax probability of the predicted label is greater. In effect, the unsupervised component is mostly zero towards the beginning of training and semi-supervised learning only begins in later batches as the proportion of confident predictions increases. This has the advantage of being based on a metric that attempts to capture the quality

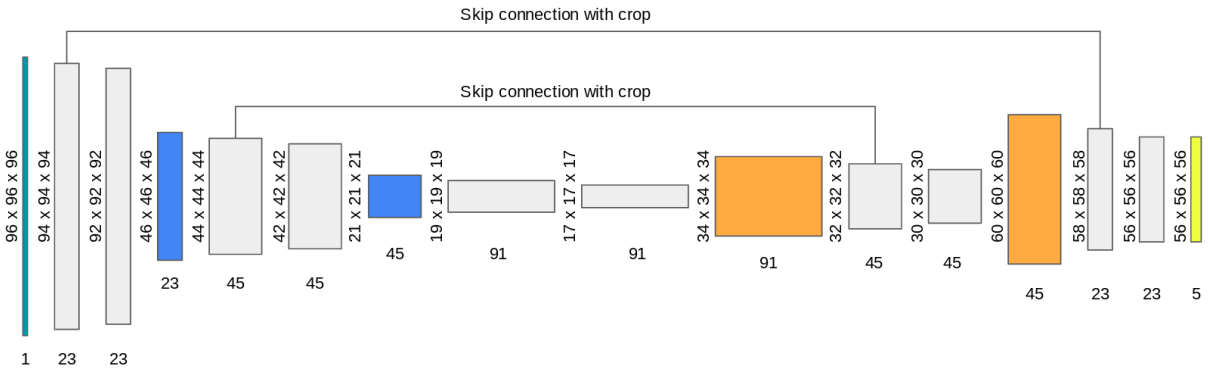


Figure 2: The 3D UNet-like network architecture: in teal is the input, in grey are 3D convolutions with stride 1 and no padding, in blue are downsampling convolutions with stride 2, in orange are plain upsampling layers with scale factor 2 and in yellow is the output. The numbers on the side of each layer are dimensions of the feature maps, while the numbers along the bottom show the number of feature maps.

of the predictions, but its adaptation for segmentation is less straight forward. Segmentation is akin to a joint classification task for each voxel (3D pixels), so that each will have its own set of softmax probabilities, hence a confidence threshold will have to be applied on a voxel level. Further, voxels close to the segmentation boundaries are the most important, but they are also typically the ones with lower confidence, meaning that they would be under represented in the loss function. Nevertheless, preliminary experiments were conducted with this setup, which did not exhibit very good convergence behaviour and was abandoned early on, hence results are not presented. Another feature that was not carried over to this work is the use of Kullback-Leibler divergence [30] as unsupervised loss, which required consistency in terms of the predicted probability distribution over all possible classes, which is a stronger condition than simply making the same prediction. This also needs to be applied on a voxel level, which did not transfer well.

3.2 Architecture

The network architecture, as shown in Fig. 2, resembles a fully-convolutional [31] 3D Unet [32]. The network is organised into a down-sampling and an up-sampling phase, with skip connections between the corresponding blocks in each phase, thus resembling an autoencoder [33] with skip connections. The convolutional layers use $3 \times 3 \times 3$ kernels, followed by batch normalisation [34] and leaky ReLU activation [35]. There are two downsampling convolutional layers with strides of two, which are mirrored by two unsampling layers that doubles the size of each dimension and linearly interpolate the intermediate values. The number of filters approximately doubles after each downsampling layer and roughly halves after each upsampling layer. There are skip connections between corresponding downsampling and upsampling blocks, where the earlier output is center-cropped to the correct dimensions. The output layer is a softmax-activated convolutional layer with a $1 \times 1 \times 1$ kernel, which is equivalent to a fully connected layer with shared weights; this is the only convolutional layer that does not use batch normalisation. The advantage of a fully convolutional setup is that a trained network can make predictions on inputs of different dimensions, which is crucial for the augmentation procedure as explained in Section 3.3. The output dimensions are smaller than the input, so a segmentation is only pro-

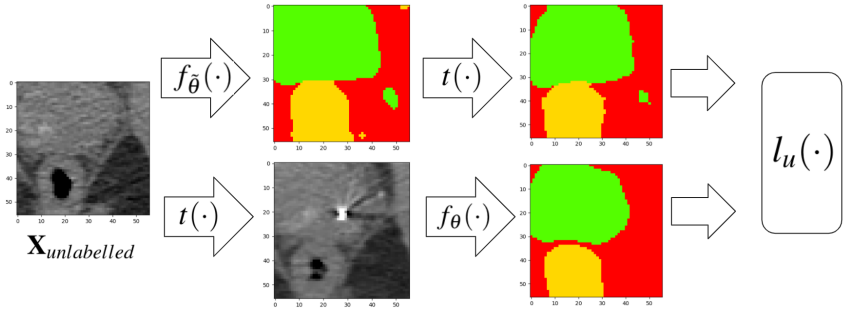


Figure 3: The unsupervised component, where the upper branch shows an unlabelled image being segmented then transformed, while the lower branch shows the same image being transformed then segmented. l_u is the unsupervised loss function.

duced for the center of the input image. A figure of the network is available in the supplement.

3.3 Training Procedure

This section describes the loss function, which follows the same structure as existing literature and Stochastic Weight Averaging (SWA), which we introduce for semi-supervised medical image segmentation. The pseudo codes are provided in Algorithms 1 and 2.

3.3.1 Loss Function

The loss function is a weighted sum of supervised and unsupervised losses, as shown in Eq. (1). The unsupervised loss is based on transformation consistency, where the segmentation of a spatially transformed image should be the same as the transformed segmentation of the original image. Let f represents the forward pass of the network and t represents a transformation, we assert that $f \circ t = t \circ f$. This is illustrated in Fig. 3, where any deviation between $t \circ f_{\tilde{\theta}}(\mathbf{X})$ and $f_{\theta} \circ t(\mathbf{X})$ are penalised. $\tilde{\theta}$ denotes non-trainable parameters, so $t \circ f_{\tilde{\theta}}(\mathbf{X})$ is used as a "fake" label. This setup follows Li *et al.* [5] but differs from Bortsova *et al.* [4], who used a Siamese model structure. t is a random choice between affine transformation (scaling and shearing) and elastic deformation. The inputs to the network are sampled patches from the full CT scans, but a larger patch is used as input to t and then cropped, so as to ensure a smooth transformation along image boundaries. This also means that $f_{\tilde{\theta}}(\mathbf{X})$ takes a larger input window, as it is the input to t; this is made possible by the fully convolutional setup. The loss function is defined as follows:

$$L(X_l, X_u) = \mathop{\mathbb{E}}_{(\mathbf{X},\mathbf{Y})\in X_l} l_s(f_{\theta}(\mathbf{X}), \mathbf{Y}) + \lambda \mathop{\mathbb{E}}_{\mathbf{X}\in X_u} l_u(f_{\theta} \circ t(\mathbf{X}), t \circ f_{\tilde{\theta}}(\mathbf{X})), \tag{1}$$

where X_l and X_u are the labelled and unlabelled data sets, $\mathbf{X}$ and $\mathbf{Y}$ are tensor representations of an input image and its label, l_s and l_u are the supervised and unsupervised loss functions, both of which are Dice loss in this paper, and λ is a weighting coefficient.

3.3.2 Dice Loss

Dice loss is based on the Sørensen–Dice coefficient (DSC):

$$DSC = \frac{2|A \cap B|}{|A| + |B|} \tag{2}$$

which measures the proportion of overlap between two sets. This is relevant because image segmentation is essentially a voxel (3D pixel) by voxel classification task, so A could represent the set of voxels that are predicted as the prostate and B those whose label is the prostate. Since DSC is a coefficient between 0 and 1, where 1 represents perfect overlap, 1 - DSC is used for minimisation.

A few adaptations are made to define dice loss, which is as follows:

$$DSC_k = \frac{2\sum_{elements} \hat{\mathbf{Y}}_k \circ \mathbf{Y}_k}{\sum_{elements}(\hat{\mathbf{Y}}_k + \mathbf{Y}_k)} \tag{3a}$$

$$DiceLoss = 1 - \frac{\sum_{k=1}^{K} DSC_k}{K} \tag{3b}$$

where $\hat{\mathbf{Y}}_k$ is a $B \times L \times W \times D$ tensor containing the predicted probabilities of each voxel for class k and $\mathbf{Y}_k$ is the binary indicator tensor of the label for class k (i.e. the k-th slice of one-hot encoded labels), K is the number of classes, B is the batch size, $L \times W \times D$ is the dimension of the input, $\sum_{element}$ is the sum of tensor elements and $\circ$ is the Hadamard (element-wise) multiplication. In the numerator of Eq. (3a), the A from Eq. (2) is replaced by a tensor containing the predicted probabilities that each voxel is the prostate and B is a binary indicator tensor of which voxels are labelled as such, the two are multiplied element-wise and then summed. This ensures a smooth progression in the loss function as the predictions become more confident. It is therefore similar to calculating a cross-entropy loss for each voxel, except probabilities are used here instead of log-probabilities. Secondly, note that $\hat{\mathbf{Y}}_k$ and $\mathbf{Y}_k$ are defined for the whole batch, which is not the same as calculating a loss for each input then averaging across the batch, but it is more efficient to implement this way and its numerical properties are the same. Finally, the set cardinality operator is implemented as a sum of array elements. This calculation is repeated for each organ and averaged to get the final loss.

3.3.3 Stochastic Weight Averaging

SWA was first introduced by Izmailov *et al.* [10], who proposed to average over a sample of network weights during the latter stages of Stochastic Gradient Descent (SGD). This is beneficial because the loss surface traversed by SGD near the end of training is approximately convex [36], which implies that the average of several solutions is likely to have a lower loss. Athiwaratkun *et al.* [29] examined semi-supervised learning and consistency loss specifically, albeit in a classification setting, finding that the sampled points are further away from each other in Euclidean space, hence that SGD traverses a wider region of the weight space; from this, they concluded that averaging is especially beneficial for semi-supervised learning.

In our implementation, 10 sets of weights are sampled after the network has reached convergence, then they are averaged to produce the SWA model. The only inputs to producing an SWA model are the weights produced during the normal course of training, then one additional forward pass is required on the training data to recalibrate any batch normalisation layers; hence the additional cost of producing such a model is essentially zero.

3.3.4 Dealing with 3D Inputs

Working with sampled windows of $96 \times 96 \times 96$ from the full CT scans on an RTX6000 GPU with 24GB of memory, the largest possible batch size was 14. In order to support the large ratio of unlabelled to labelled images in the training set, which is also reflected in the composition of each batch, gradients are backpropagated for every sub-batch of six windows, while the optimiser only makes a step after a full batch. Each full batch always consists of one labelled sub-batch and multiple unlabelled sub-batches, which reflects the ratio of unlabelled to labelled images in the training set. Consequently, the unsupervised loss for each sub-batch is downscaled by a factor of r, the ratio of unlabelled to labelled input, otherwise the weight of the unsupervised component would implicitly increase; this is shown in line 27 of Algorithm 1. Mathematically, the loss is now an arithmetic mean over the sub-batches, which has the same expected value as the loss calculated on the full batch.

Algorithm 1 Training Procedure

1: **procedure** TRAIN(X_l, X_u, f_θ) ▷ subscript l means labelled and u unlabelled
2: **for** $e \leftarrow 1, num\_epochs$ **do**
3: **for all** labelled batch $\mathbf{B}_l$ in X_l **do**
4: LABELLEDITER($\mathbf{B}_l$)
5: **if** $e \geq start\_ssl$ **then**
6: UNLABELLEDITERS(X_u)
7: **end if**
8: **end for**
9: $\theta \leftarrow$ OPTIMISERSTEP
10: **if** $e \geq start\_swa$ **then**
11: SAVEPARAMETERS(θ)
12: **end if**
13: **end for**
14: **end procedure**

15: **procedure** LABELLEDITER($\mathbf{B}_l$)
16: $\hat{\mathbf{Y}}_l \leftarrow f_\theta(\mathbf{B}_l)$
17: $SupLoss \leftarrow$ DICELOSS($\hat{\mathbf{Y}}_l, \mathbf{Y}_l$) ▷ compute supervised loss
18: BACKPROPAGATE($SupLoss$)
19: **end procedure**

20: **procedure** UNLABELLEDITERS(X_u)
21: **for** $i \leftarrow 1, r$ **do** ▷ r is the ratio of unlabelled to labelled inputs
22: $\mathbf{B}_u \leftarrow$ NEXTBATCH(X_u)
23: $\mathbf{Y}_u \leftarrow$ TRANSFORM($f_{\tilde{\theta}}(\mathbf{B}_u)$)
24: $\mathbf{Y}_u \leftarrow \arg\max_k \mathbf{Y}_u$ ▷ probabilities are converted to class labels
25: $\hat{\mathbf{Y}}_u \leftarrow f_\theta($TRANSFORM$(\mathbf{B}_u))$
26: $UnsupLoss \leftarrow$ DICELOSS($\hat{\mathbf{Y}}_u, \mathbf{Y}_u$) ▷ compute unsupervised loss
27: $UnsupLoss \leftarrow \frac{\lambda}{r} UnsupLoss$ ▷ see Section 3.3.4
28: BACKPROPAGATE($UnsupLoss$)
29: **end for**
30: **end procedure**

Algorithm 2 Generate SWA Model

function DOSWA($f, saved\_parameters, X_l$)
    $\theta \leftarrow$ MEAN($saved\_parameters$)
    SETBATCHNORM(f_θ, X_l) ▷ Do one forward pass to calibrate BatchNorm layers
    return f_θ
end function

3.4 Types of Transformation

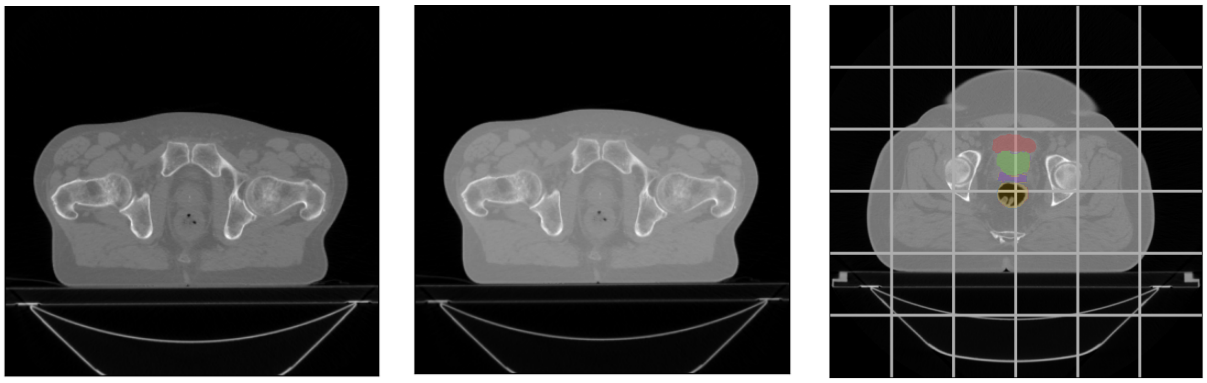
Three types of transformation were investigated: voxel intensity shift, affine transformations and elastic deformations; examples are shown in Figs. 4 and 5. Intensity shift was implemented as a uniform shift to the voxel values in the region excluding the background. Affine transformations included random scaling and shearing in each of the three input dimensions. Elastic deformation was implemented with the code from [1], which randomly perturbs a sample of control point in the input image, then interpolates the location of the voxels between these points, resulting in a smooth non-linear transformation. In line with the conclusion in [16] that the appropriateness of the transformation has a big influence on network performance, Section 4.3 compares the relative effectiveness of each.

An implementation detail to note is that affine and elastic transformations are applied to a larger input image, which are then cropped to ensure a smooth transformation along the edges of the cropped output. Consequently, the input to $t(\circ)$ in Fig. 3 has a larger dimension than its output, which means that the input to $f_{\tilde{\theta}}(\circ)$ is also larger than other forward passes during training; this is only possible because f is a fully convolutional network.

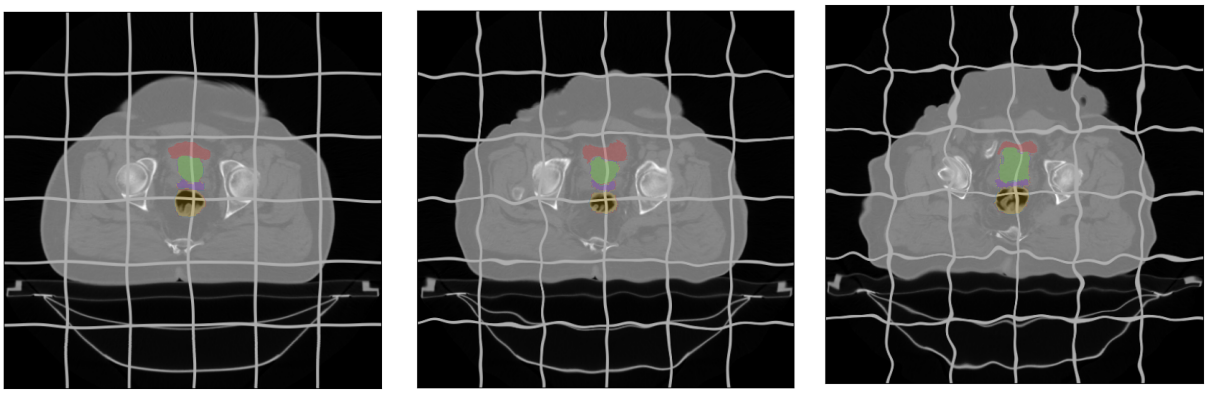
4 Experiments and Results

4.1 Data

The dataset is composed of prostate CT scans from three hospitals using different scanners. All images have a dimension of 512×512, but varying numbers of slices and voxel spacing. Leiden University Medical Center (LUMC) in the Netherlands contributes with 399 scans, which have 68-240 slices and a voxel size of $1.0 \times 1.0 \times 3.0mm$. The second dataset is from Haukeland Medical Center (HMC) in Norway and has 161 scans, which are composed of 91-218 slices and a voxel size of $0.9 \times 0.9 \times 1.5mm$. The last dataset comes from Erasmus Medical Center (EMC) in the Netherlands and has 42 scans with 90 - 180 slices and a voxel size of $0.9 \times 0.9 \times 2-3mm$. This is summarised by Table 1. Four target classes are delineated in all images, the prostate, seminal vesicles, bladder and rectum, and these were done manually by radiation oncologists. The voxel intensities were clipped to remove extreme values, then normalised to a range of -1 to 1. During training and validation, a class-balanced sampler was used to sample three windows of dimension $96 \times 96 \times 96$ online from each image, which are used as inputs to the network. The network weights with the lowest validation loss are used for inference on the test set, for which a sliding window sampler is used.



(a) Intensity Down 0.2 (b) Intensity Up 0.2 (c) No transformation



(d) Elastic Small (e) Elastic Medium (f) Elastic Large

Figure 4: Examples of the intensity and elastic deformation transformations. For intensity, the background is excluded and shifts of up to ± 0.2 were applied to original voxel values in the range $[-1, 1]$. For elastic deformation, three different magnitudes were used. The grid was added for illustrative purpose, to show the effect of the transformation. Organ labels were also overlaid in colour to show the effect on each target class. The images shown are cross-sectional slices of the full 3D scan.

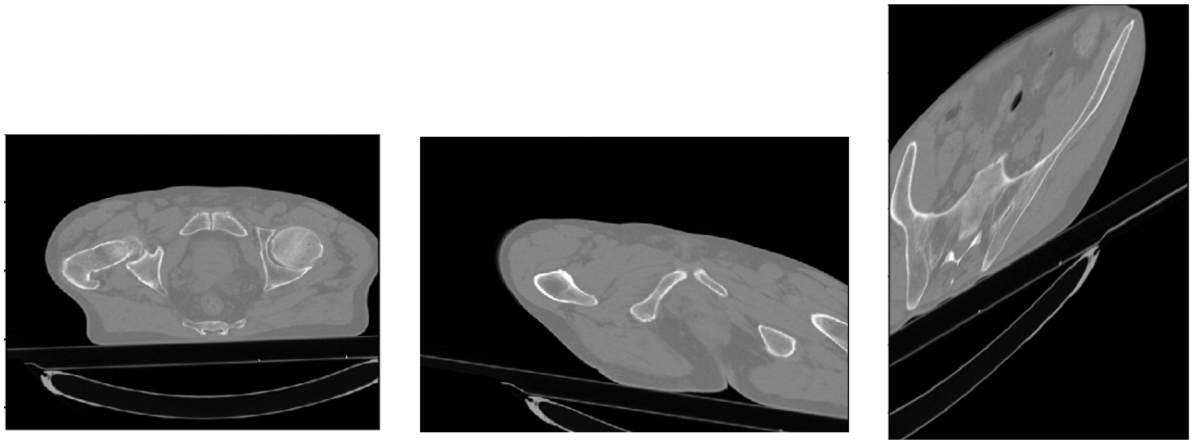


Figure 5: Examples of the three magnitudes of affine transformation used: small, medium, and large, which consisted of random scaling and shearing in each dimension. The images shown are cross-sectional slices of the full 3D scan.

Table 1: Composition of the dataset and technical attributes of the images produced by different scanners. The sources are Leiden University Medical Center, Erasmus Medical Center in Rotterdam and Haukeland Medical Center in Bergen.

Source	LUMC	EMC	HMC
Number of scans	399	42	161
Image Dimension	512 x 512 x (68 – 240 slices)	512 x 512 x (91 – 218 slices)	512 x 512 x (90 – 180 slices)
Voxel Spacing	∼1.0 x 1.0 x 3.0 mm	∼0.9 x 0.9 x 1.5 mm	∼0.9 x 0.9 x 2-3 mm

4.2 Experimental Setup

Two different experiments are presented in this paper. For each experiment, results are presented for a fully-supervised baseline and semi-supervised networks, both of which use the same UNet-like [32] architecture as described in Section 3.2. The difference is that λ in Eq. (1) is set to 0 for the baseline, while for SSL it is initially 0 for a supervised phase, then 0.5 for a semi-supervised phase. This was so that the unlabelled predictions could reach a reasonable accuracy for consistency loss to work.

For the experiment on the transformation procedure (Section 4.3), the supervised baseline was trained for 500 epochs over a labelled set of 105 CT scans. The validation set consisted of 37 full scans, the test set 53, and another 105 were treated as unlabelled. Each semi-supervised network was trained for another 300 epochs using the weights from the supervised baseline as the starting point. The RAdam optimiser [37] was used with an annealing stepped learning rate. The aim of this experiment is to determine the optimal transformation procedure for SSL.

For the experiment with SWA (Section 4.4), results are presented for high and low labelled data regimes. For the high data regime, 105 CT scans were randomly selected as the labelled training set, while 407 scans were treated as unlabelled, 37 scans as validation and 53 scans as test. This was repeated three times to generate three folds for the high data regime and two folds for the low data regime. Each source hospital is represented in the same proportion in each split. For the low data regime, the labelled training set was reduced to 20 CT scans. In Section 4.4, experiments for the high data regime are labelled “105” and low data regime are labelled “20”.

For the high data regime, the Base network was trained for 400 epochs, while the SSL model had 100 epochs of supervised training, plus 300 epochs of semi-supervised training. Both models used a random start, in which all convolutional layers were initialised from a random normal distribution of $\mathcal{N}(0, 0.02^2)$. The SWA models consisted of averaged weights over the last 50 epochs, sampled every 5 epochs. For the low data regime, the networks were trained for a total of 1110 epochs and the semi-supervised phase started after 400 epochs. The SWA weights were averaged over the last 100 epochs, sampled every 10 epochs; the lower sample rate tries to account for the fact that each epoch consists of fewer iterations. Finally, all networks were trained with the RAdam optimiser [37] with a constant maximum learning rate of 10^{-4}.

We also compared the proposed SWA approach to three state-of-the-art methods in abdominal CT radiotherapy: Cross-Stitch [38] is a deep learning approach that shares weights between a segmentation and registration CNN, Elastix [39] is a conventional iterative registration method

and a Hybrid model [40] that feeds CNN segmentations of the bladder to an iterative approach as prior knowledge.

Three different metrics were calculated: Sørensen–Dice coefficient (DSC), mean surface distance (MSD) and 95% Hausdorff distance (HD). DSC is defined in (2). MSD is the mean distance between voxels on the segmentation surface and the closest surface voxel in the ground truth, while HD is an outlier measure of the greatest distance between segmentation boundaries, which in this case is the 95th-percentile of the surface distances. While DSC can be conveniently implemented as a loss function (see Section 3.3), it measures the percentage of common voxels in the prediction and label, which means that it tends to be higher for larger organs. MSD overcomes this disadvantage by measuring the mean deviation between the segmentation boundaries in millimeters, which is also more meaningful in determining the clinical usefulness and safety of the methods. Furthermore, having good segmentation on average is not enough to guarantee safety, since excess radiation on any healthy tissue could lead to complications, hence the inclusion of HD. The Wilcoxon signed-rank test [41] is used to test for statistical significance against the baseline, because it is a non-parametric test and the evaluated metrics on the test set do not follow a normal distribution; specifically, H_0 is that the median performance on the test set is equal between two methods.

4.3 Transformation Procedure

Table 2: Test set results for different transformation procedures. Lower values are better. † denotes a difference from the baseline at 5% statistical significance using a Wilcoxon signed rank test.

(a) MSD (mm)

	Prostate		Seminal vesicles		Rectum		Bladder	
	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median
Baseline	1.87 ± 1.7	**1.51**	3.19 ± 9.1	1.97	2.28 ± 1.5	1.71	0.92 ± 0.6	0.74
Intensity	$1.88 \pm 1.1^{\dagger}$	1.67	2.50 ± 3.7	1.89	$2.56 \pm 1.4^{\dagger}$	2.01	$1.14 \pm 0.9^{\dagger}$	0.84
Affine S	$1.79 \pm 0.7^{\dagger}$	1.64	2.50 ± 4.3	1.76	2.35 ± 1.5	1.77	0.99 ± 0.7	0.79
Elastic S	$1.77 \pm 0.8^{\dagger}$	1.53	2.45 ± 3.9	1.91	$2.55 \pm 1.7^{\dagger}$	1.90	0.97 ± 0.7	0.75
Aff+Ela M	$\mathbf{1.71 \pm 0.7}$	1.55	$\mathbf{2.32 \pm 3.4}$	1.73	2.26 ± 1.5	1.67	0.85 ± 0.6	0.69
Aff+Ela L	1.72 ± 0.7	1.55	2.49 ± 4.7	**1.70**	$\mathbf{2.19 \pm 1.5}$	**1.64**	$\mathbf{0.80 \pm 0.5}^{\dagger}$	**0.65**

(b) 95%HD (mm)

	Prostate		Seminal vesicles		Rectum		Bladder	
	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median
Baseline	5.9 ± 4.3	5.7	9.2 ± 10.4	7.3	$\mathbf{13.9 \pm 9.7}$	10.3	4.3 ± 3.7	**3.0**
Intensity	6.0 ± 2.9	6.0	9.0 ± 8.0	6.9	15.4 ± 9.8	12.0	$5.7 \pm 6.2^{\dagger}$	3.2
Affine S	5.9 ± 2.7	6.0	8.7 ± 7.8	7.2	14.6 ± 10.2	10.8	5.2 ± 6.0	**3.0**
Elastic S	5.7 ± 2.8	5.4	8.7 ± 7.7	7.7	15.7 ± 11.5	10.1	4.6 ± 5.3	**3.0**
Aff+Ela M	5.6 ± 2.4	5.2	$\mathbf{8.6 \pm 8.1}$	6.4	14.1 ± 10.5	**9.6**	4.0 ± 3.9	**3.0**
Aff+Ela L	$\mathbf{5.5 \pm 2.6}$	**4.9**	8.9 ± 9.7	**6.0**	14.1 ± 10.1	10.0	$\mathbf{3.9 \pm 3.9}^{\dagger}$	**3.0**

In order to optimise the function $t(\cdot)$, three different types of transformation were used, as

Table 3: Relative performance by data source, lower values are better. The test set was divided into three groups corresponding to which hospital provided the image, for which group means and medians were calculated for each evaluation metric and organ, these are then ranked from one to three and the summed ranks are presented below. Aff+Ela L from Table 2 is used to represent SSL and "Base" refers to the supervised baseline.

	SSL Mean	SSL Median	Base Mean	Base Median
EMC	34	26	35	27
HMC	**14**	**20**	**15**	**19**
LUMC	24	26	22	26

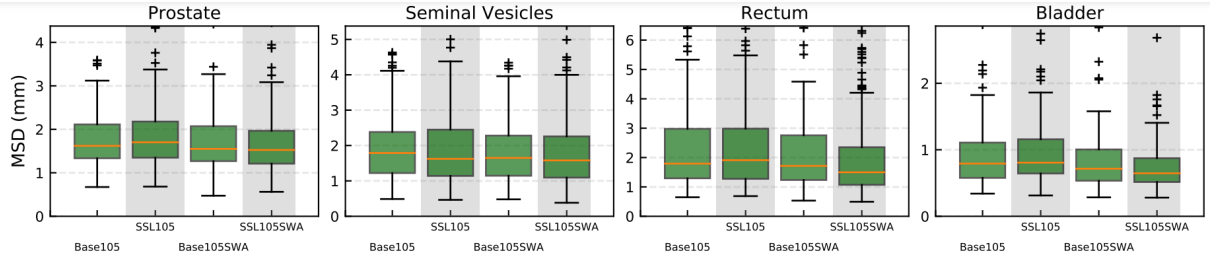


Figure 6: Distribution of test set MSD values for the high data regime.

discussed in Section 3.4, as well as different magnitudes of transformation, as shown in Figs. 4 and 5. A selection of results are presented in Table 2, where the $\mu + \sigma$ column gives a sense of variance, while the median gives a sense of the average without outliers. The † denotes that the evaluated metric for the test set as a whole is statistically significantly different to the baseline, using a 95% confidence interval on a Wilcoxon signed-rank test [41], it is not specifically related to the $\mu + \sigma$ column.

Intensity gave consistently worse results than Elastic S and Affine S, except for 95%HD on the seminal vesicles. It was also often statistically significantly worse than the baseline. Compared to the baseline, Affine S and Elastic S also did not achieve better results, but most metrics did improve when the magnitude of the transformations was increased. Aff+Ela M and Aff+Ela L used a mix of affine and elastic transformations with increasing magnitudes of transformations, which produced better mean MSD but not the median; 95%HD was also better. However, the lack of statistical significance on all organs except the bladder indicates that the improvements are small and that the conclusions could change under a different data split.

Lastly, Table 3 shows how performance differs for each source hospital. The test set was split into three group corresponding to each source hospital, then each group was ranked by their median and mean for each evaluation metric and organ combination; the summed ranks are presented in Table 3, for which lower values are better. HMC consistently gave the best results, followed by LUMC and then EMC.

4.4 Stochastic Weight Averaging

Table 4 shows that in the high data regime, SSL105 performed similarly to Base105, with sizeable improvement for the seminal vesicles, but statistically significantly worse on the prostate,

Table 4: Test set MSD (mm) values for the high and low data regimes. Lower values are better. † signifies 5% statistical significance vs Base20 and Base105 using a Wilcoxon Signed-Rank Test.

	Prostate		Seminal vesicles		Rectum		Bladder	
	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median
Base105	1.85 ± 1.1	1.63	2.24 ± 2.4	1.74	2.43 ± 1.7	1.81	1.04 ± 0.8	0.80
SSL105	$1.84 \pm 0.7^{\dagger}$	1.72	2.20 ± 2.4	1.63	2.44 ± 1.7	1.86	1.27 ± 1.5	0.80
Base105SWA	$1.76 \pm 1.0^{\dagger}$	1.55	$2.10 \pm 2.1^{\dagger}$	1.63	$2.32 \pm 1.8^{\dagger}$	1.72	$1.09 \pm 1.5^{\dagger}$	0.73
SSL105SWA *(Proposed)*	$\mathbf{1.64} \pm 0.6^{\dagger}$	**1.51**	$\mathbf{2.08} \pm 2.3^{\dagger}$	**1.60**	$\mathbf{2.07} \pm 1.5^{\dagger}$	**1.49**	$\mathbf{0.86} \pm 0.9^{\dagger}$	**0.66**
Base20	2.41 ± 1.4	2.07	4.12 ± 7.9	2.26	3.40 ± 2.7	2.48	1.94 ± 3.7	0.97
SSL20	$2.12 \pm 0.8^{\dagger}$	1.97	2.57 ± 1.6	2.17	3.56 ± 3.1	2.33	1.54 ± 1.6	0.92
Base20SWA	$2.63 \pm 1.9^{\dagger}$	2.15	4.04 ± 5.1	2.29	3.71 ± 3.0	2.69	1.99 ± 4.2	0.99
SSL20SWA *(Proposed)*	$\mathbf{1.94} \pm 0.8^{\dagger}$	**1.77**	$\mathbf{2.46} \pm 2.6^{\dagger}$	**1.79**	$\mathbf{2.81} \pm 2.0^{\dagger}$	**2.27**	$\mathbf{1.30} \pm 1.4^{\dagger}$	**0.80**

Table 5: Test set MSD (mm) comparison against other state-of-the-art methods.

	Prostate		Seminal vesicles		Rectum		Bladder	
	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median
SSL105SWA *(proposed)*	1.64 ± 0.6	1.51	2.08 ± 2.3	1.60	$\mathbf{2.07} \pm 1.5$	**1.49**	$\mathbf{0.86} \pm 0.9$	**0.66**
Cross-Stitch *Segmentation* [38]	1.88 ± 2.2	1.21	4.73 ± 8.0	1.42	3.61 ± 5.0	2.18	2.45 ± 2.4	1.24
Elastix [39]	$\mathbf{1.42} \pm 0.7$	**1.17**	2.07 ± 2.6	1.24	3.20 ± 1.6	3.07	5.30 ± 5.1	3.27
Hybrid [40]	1.55 ± 0.6	1.36	$\mathbf{1.65} \pm 1.3$	**1.22**	2.65 ± 1.6	2.36	3.81 ± 3.6	2.26

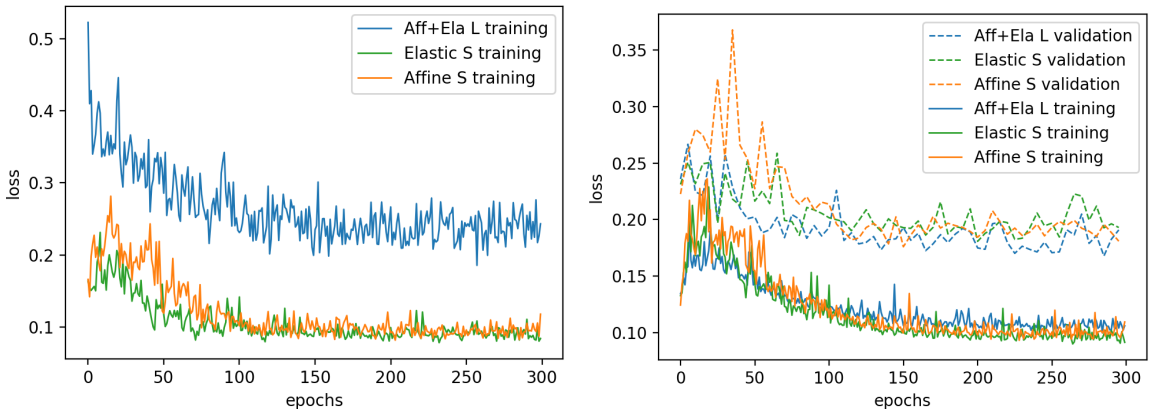
while the rectum and bladder were similar. Both SWA models, supervised and semi-supervised, achieved statistical significance in their outperformance over Base105, while SSL105SWA also reached statistical significance over Base105SWA for the rectum and bladder. Moreover, SSL105SWA produced more consistent predictions, as shown by the distribution of the test set MSD in Fig. 6. Specifically, SSL105SWA shows tighter and lower interquartile ranges for the prostate, rectum and bladder, as well as the lowest quartiles for the seminal vesicles. Hence the combination of SSL and SWA produced the best segmentation performance as well as the lowest variance.

In the low data regime, SSL20 showed more marked improvements against Base20 than in the high data regime, although it is not statistically significant for all organs. Similarly to the high data regime, SSL20SWA produced the best results and they are also statistically significant. However, Base20SWA did not improve over Base20 in the low data regime.

Table 5 provides a comparison in the high data regime against existing state-of-the-art results, which are fully supervised image registration methods. The performance of our pure segmentation method falls short for the prostate and seminal vesicles, but outperformed substantially for the rectum and bladder. Our proposed model takes approximately 0.5 seconds to segment a full CT scan, which is comparable to Cross-Stitch and shorter by an order of minutes than the iterative and hybrid methods [38].

5 Discussion

The experiment in Section 4.3 set out to determine the optimal transformation procedure for SSL. The results indicate that intensity shift is less useful; a possible explanation could be



(a) Unsupervised Loss

(b) Supervised Loss

Figure 7: Learning curves for the supervised and unsupervised components of the loss function for a selection of experiments. Fig. 7a shows that network predictions are much more consistent for less extreme transformations, which might not provide as much new information to learn. Fig. 7b shows that more extreme transformations had a bigger regularisation effect and improved generalisation, because the training loss was higher but validation loss was lower.

that it does not actually affect the segmentation boundary, whereas the other transformations are more geometric in nature, thus more similar to the task at hand. Another possible reason is more specific to CT scans, in which intensity values do not vary much between different patients and should fall within a set range for different types of tissue. Consequently, intensity was dropped from further experiments. The results for Aff+Ela M and Aff+Ela L in Table 2 suggests that SSL has led to more consistent predictions by reducing the extremity of the outliers, since there were improvements in the mean MSD and 95%HD, but not in median MSD.

Despite the lack of statistical significance, the general improvement when larger transformations were used is surprising. The S versions of the transformations, which gave worse results, were tuned so that the resulting images are as similar as possible to the original dataset, while still introducing some variations. Fig. 7a shows that a possible explanation could be that the predictions for less extreme transformations were already very consistent, therefore it had very little new information to learn, while more extreme transformations beget more mistakes, as shown by the higher unsupervised loss. Another interesting observation from Fig. 7b is that Aff+Ela L had higher training loss but lower validation loss, which means that more extreme transformations had a similar effect to regularisation, improving generalisation. A final observation is that only the bladder showed statistically significantly better results; the explanation could be that more extreme transformations are more appropriate for the bladder, as it is a larger organ that shows more variation from day to day.

Table 3 shows that there are systematic differences between the source hospitals to which the network has not been able to fully adapt, which suggests that improved preprocessing might yield better results. However, not all differences can be eliminated, since the quality of the segmentation labels also depends on the differing aptitudes of the clinicians who produced them. Furthermore, outliers appear to be a large driver of this discrepancy, since the scores are more dispersed for the mean.

The experiment in Section 4.4 set out to investigate the effectiveness of SWA for transformation-consistent SSL in an image segmentation application, as well as the relative effectiveness of

the model in high and low labelled data regimes. Moreover, a comparison to previous state-of-the-art results on prostate CT datasets was also provided. The results for SSL without SWA, SSL20 and SSL105, are in line with existing literature [4, 5], which showed that SSL outperforms supervised learning when a small labelled dataset is used, but performs similarly when a large labelled training set is used. The significant finding of this paper is that the use of SWA with SSL leads to further improvement that is also statistically significant in both the high and low data regimes. Moreover, this is achievable without additional training cost, as discussed in Section 3.3.

To differentiate how much of the improvement is attributable to SSL and SWA respectively, supervised SWA models (Base20SWA and Base105SWA) were also provided for a second comparison. Our results showed that SWA is especially beneficial for SSL, but can also be beneficial for supervised learning in the high data regime. This could indicate that SWA works well whenever a large amount of training data is available, both labelled and unlabelled. In the high data regime, our proposed method show the most marked outperformance on the rectum and bladder, which is consistent with the experiment on the transformation procedure. In the low data regime, the outperformance was also substantial on the prostate and seminal vesicles, which could reflect the supervised baseline being deprived of training data.

Compared to the current state-of-the-art, our proposed method also outperformed on the rectum and bladder, but underperformed on the prostate and seminal vesicles. A likely reason is that the other methods all use registration to some extent, meaning that the segmentation from a planning scan is available as an input. This is particularly helpful for the prostate and seminal vesicles, which show little spatial variation from day to day, but less so for the rectum and bladder, which vary a lot. Our proposed method do not make use of any prior segmentation, so it is a pure segmentation method. A caveat, however, is that the comparison to the other methods is not direct, since the datasets used are not exactly the same, albeit with significant overlap.

5.1 Scope for Improvement

While these are promising results, there are several areas for improvement. Firstly, more parameter tuning could be helpful; for example the maximum learning rate in Section 4.4 was carried over from previous research on the same data set [38], as were many other network parameters, which could benefit from more specific tuning for SSL. Secondly, more detailed study of the effect of different types and magnitudes of transformation on each organ could be helpful, as Section 4.3 suggests that larger magnitudes might lead to better performance on the bladder, but this could be confirmed with more certainty. Moreover, Affine L introduces padded regions into the transformed image, as shown in Fig. 8; presently it is unclear whether this is helpful or detrimental. Lastly, it would be interesting to see how SSL with SWA performs on a publicly available dataset.

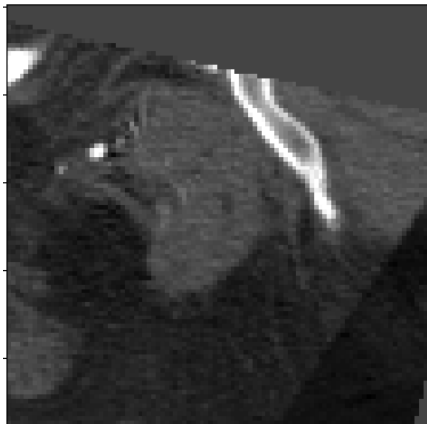


Figure 8: An example of a sampled window that has been transformed by Affine L, showing the ingress of padded regions along the top and right edges.

5.2 Conclusion

This paper set up to determine if semi-supervised methods can improve on a fully supervised baseline in a CT image segmentation task. In contrast to previous applications of transformation consistent SSL, the use of SWA is novel in our domain of application. The results showed that SWA was beneficial in both high and low labelled data regimes, but can also be useful for supervised learning when a large amount of training data is available. The improvement in the high data regime is particularly significant, since existing research showed that SSL do not outperform supervised learning in this setting without SWA. Further, since this gain in performance comes with no additional training cost, SWA should be adopted as a matter of course for transformation-consistent semi-supervised methods. In addition, we also found that larger-than-realistic transformations can be beneficial, especially for organs that have more day-to-day variation.

A Appendix

A.1 Additional Results Tables

Table A1: Test set Dice values for the high data regime. Higher values are better. † signifies 5% statistical significance vs Base105 using a Wilcoxon Signed-Rank Test.

	Prostate		Seminal vesicles		Rectum		Bladder	
Output Path	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median
Base105	0.83 ± 0.08	0.85	0.64 ± 0.15	0.67	0.83 ± 0.08	0.85	0.93 ± 0.06	0.95
SSL105	0.84 ± 0.05	0.84	0.66 ± 0.14	0.68	0.83 ± 0.07	0.85	0.92 ± 0.07	0.94
Base105SWA	$0.84 \pm 0.07^{\dagger}$	**0.86**	$0.65 \pm 0.15^{\dagger}$	0.68	$0.84 \pm 0.08^{\dagger}$	0.85	$0.93 \pm 0.08^{\dagger}$	0.95
SSL105SWA	$\mathbf{0.85 \pm 0.04}^{\dagger}$	0.86	$\mathbf{0.67 \pm 0.13}^{\dagger}$	**0.69**	$\mathbf{0.86 \pm 0.06}^{\dagger}$	**0.88**	$\mathbf{0.94 \pm 0.05}^{\dagger}$	**0.95**

References

[1] Chao Huang, Hu Han, Qingsong Yao, Shankuan Zhu, and S. Kevin Zhou. 3D U^2-net: A 3D universal U-net for multi-domain medical image segmentation. In *Medical Image Com-*

Table A2: Test set 95%HD values for the high data regime. Lower values are better. † signifies 5% statistical significance vs Base105 using a Wilcoxon Signed-Rank Test.

	Prostate		Seminal vesicles		Rectum		Bladder	
Output Path	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median
Base105	6.0 ± 3.3	5.7	8.3 ± 6.2	6.9	14.6 ± 11.3	10.4	5.0 ± 5.0	3.0
SSL105	6.2 ± 2.7	5.6	8.5 ± 6.5	6.5	15.1 ± 10.6	11.4	6.5 ± 8.7	3.2
Base105SWA	$5.8 \pm 3.3^{\dagger}$	5.3	$\mathbf{7.7 \pm 5.2}^{\dagger}$	6.3	14.1 ± 10.4	11.0	$5.2 \pm 7.2^{\dagger}$	**3.0**
SSL105SWA	$\mathbf{5.4 \pm 2.2}^{\dagger}$	**5.2**	8.2 ± 6.6	**6.2**	$\mathbf{13.3 \pm 10.8}^{\dagger}$	**9.1**	$\mathbf{4.5 \pm 5.8}^{\dagger}$	**3.0**

Table A3: Test set Dice values for the low data regime. Higher values are better. † signifies 5% statistical significance vs Base105 using a Wilcoxon Signed-Rank Test.

	Prostate		Seminal vesicles		Rectum		Bladder	
Output Path	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median
Base20	0.78 ± 0.12	0.81	0.53 ± 0.20	0.57	0.74 ± 0.12	0.79	0.90 ± 0.13	0.94
SSL20	$0.81 \pm 0.06^{\dagger}$	0.82	$0.60 \pm 0.15^{\dagger}$	0.64	$0.79 \pm 0.10^{\dagger}$	0.81	0.90 ± 0.09	0.93
Base20SWA	0.77 ± 0.13	0.80	0.52 ± 0.21	0.55	$0.72 \pm 0.14^{\dagger}$	0.77	0.89 ± 0.13	0.93
SSL20SWA	$\mathbf{0.82 \pm 0.06^{\dagger}}$	**0.83**	$\mathbf{0.61 \pm 0.17^{\dagger}}$	**0.66**	$\mathbf{0.80 \pm 0.10^{\dagger}}$	**0.82**	$\mathbf{0.92 \pm 0.07^{\dagger}}$	**0.94**

Table A4: Test set 95%HD values for the low data regime. Lower values are better. † signifies 5% statistical significance vs Base105 using a Wilcoxon Signed-Rank Test.

	Prostate		Seminal vesicles		Rectum		Bladder	
Output Path	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median	$\mu \pm \sigma$	Median
Base20	7.9 ± 4.1	6.8	11.6 ± 11.1	9.2	17.4 ± 11.6	13.6	8.0 ± 10.3	3.6
SSL20	7.3 ± 3.6	6.2	10.1 ± 5.8	8.6	19.8 ± 16.7	14.5	6.7 ± 7.6	3.7
Base20SWA	8.1 ± 4.8	7.0	11.9 ± 9.9	8.7	18.5 ± 12.9	15.4	7.8 ± 10.1	3.9
SSL20SWA	$\mathbf{6.5 \pm 2.5^{\dagger}}$	**6.2**	$\mathbf{8.9 \pm 6.5^{\dagger}}$	**7.3**	$\mathbf{16.0 \pm 11.2}$	**13.4**	$\mathbf{6.2 \pm 7.1^{\dagger}}$	**3.5**

puting and Computer Assisted Intervention, volume 11765 of *Lecture Notes in Computer Science*, pages 291–299, 2019.

[2] Eli Gibson, Wenqi Li, Carole Sudre, Lucas Fidon, Dzhoshkun I. Shakir, Guotai Wang, Zach Eaton-Rosen, Robert Gray, Tom Doel, Yipeng Hu, Tom Whyntie, Parashkev Nachev, Marc Modat, Dean C. Barratt, Sébastien Ourselin, M. Jorge Cardoso, and Tom Vercauteren. Niftynet: a deep-learning platform for medical imaging. *Computer Methods and Programs in Biomedicine*, 158:113 – 122, 2018.

[3] Rosa Figueroa, Qing Zeng-Treitler, Sasikiran Kandula, and Long Ngo. Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, 12:8, 2012.

[4] Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen de Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *Medical Image Computing and Computer Assisted Intervention*, page 810–818. Springer International Publishing, 2019.

[5] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, and Pheng-Ann Heng. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. In *British Machine Vision Conference*, 2018.

[6] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation consistent self-ensembling model for semi-supervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):523–534, 2021.

[7] Quande Liu, Lequan Yu, Luyang Luo, Qi Dou, and Pheng Ann Heng. Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE Transactions on Medical Imaging*, 39(11):3429–3440, 2020.

[8] Christian S. Perone and Julien Cohen-Adad. Deep semi-supervised segmentation with weight-averaged consistency targets. In *International Workshop on Deep Learning in Med-*

ical Image Analysis, volume 11045 of *Lecture Notes in Computer Science*, page 12–19, 2018.

[9] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Conference on Neural Information Processing Systems*, 2016.

[10] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence*, pages 876–885, 2018.

[11] Jan-Jakob Sonke, Marianne Aznar, and Coen Rasch. Adaptive radiotherapy for anatomical changes. *Seminars in Radiation Oncology*, 29(3):245–257, 2019.

[12] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors, 2012.

[13] Benjamin Graham. Fractional max-pooling, 2015.

[14] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning, 2017.

[15] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Conference on Neural Information Processing System*, page 1195–1204, 2017.

[16] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training, 2020.

[17] David J Miller and Hasan S. Uyar. A mixture of experts classifier with learning based on both labelled and unlabelled data. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 571–577. MIT Press, 1997.

[18] Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, pages 841–848. MIT Press, 2002.

[19] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013.

[20] Yves Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. volume 17, 1 2004.

[21] Yann Lecun, Koray Kavukcuoglu, and Clement Farabet. Convolutional networks and applications in vision. pages 253–256, 05 2010.

[22] Kevin Jarrett, Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann Lecun. What is the best multi-stage architecture for object recognition? volume 12, 09 2009.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015.

[25] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2018.

[26] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, Aaron Courville, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states, 2019.

[27] Veronika Cheplygina, Marleen de Bruijne, and Josien P.W. Pluim. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54:280–296, 2019.

[28] I. Triguero, S. García, and F. Herrera. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information Systems*, 42:245–284, 2013.

[29] Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. There are many consistent explanations of unlabeled data: Why you should average. In *International Conference on Learning Representations*, 2019.

[30] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951.

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans Pattern Anal Mach Intell*, 39(4):640–651, 2017.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241, 2015.

[33] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning Internal Representations by Error Propagation*, page 318–362. MIT Press, Cambridge, MA, USA, 1986.

[34] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of Machine Learning Research*, volume 37, pages 448–456, 2015.

[35] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning*, 2013.

[36] Ian J. Goodfellow, Oriol Vinyals, and Andrew M. Saxe. Qualitatively characterizing neural network optimization problems. In *International Conference on Learning Representations*, 2015.

[37] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*, 2020.

[38] Laurens Beljaards, Mohamed S. Elmahdy, Fons Verbeek, and Marius Staring. A cross-stitch architecture for joint registration and segmentation in adaptive radiotherapy. In *Medical Imaging with Deep Learning*, volume 121 of *Proceedings of Machine Learning Research*, pages 62 – 74, 2020.

[39] Yuchuan Qiao. *Fast optimization methods for image registration in adaptive radiation therapy*. PhD thesis, Leiden University Medical Center, 2017.

[40] Mohamed S Elmahdy, Thyrza Jagt, Roel Th Zinkstok, Yuchuan Qiao, Rahil Shahzad, Hessam Sokooti, Sahar Yousefi, Luca Incrocci, CAM Marijnen, Mischa Hoogeman, and Marius Staring. Robust contour propagation using deep learning and image registration for online adaptive proton therapy of prostate cancer. *Medical physics*, 46(8):3329–3343, 2019.

[41] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.