Transformation-Consistent Semi-Supervised Learning for Prostate CT Radiotherapy

Yichao Li[a,b], Mohamed S. Elmahdy[b], Michael S. Lew[a], and Marius Staring[b,c]

[a]Leiden Institute of Advanced Computer Science, Leiden, The Netherlands
[b]Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands
[c]Department of Radiation Oncology, Leiden University Medical Center, The Netherlands

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. 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.[1] 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.[2–6]

The idea of transformation-consistent SSL, as proposed by Sajjadi *et al.*[7] 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,[2,3] so the main contribution of this paper is to show that Stochastic Weight Averaging (SWA),[8] 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.[2,3]

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.[9]

2. RELATED WORK

Cheplygina *et al.*[10] 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.

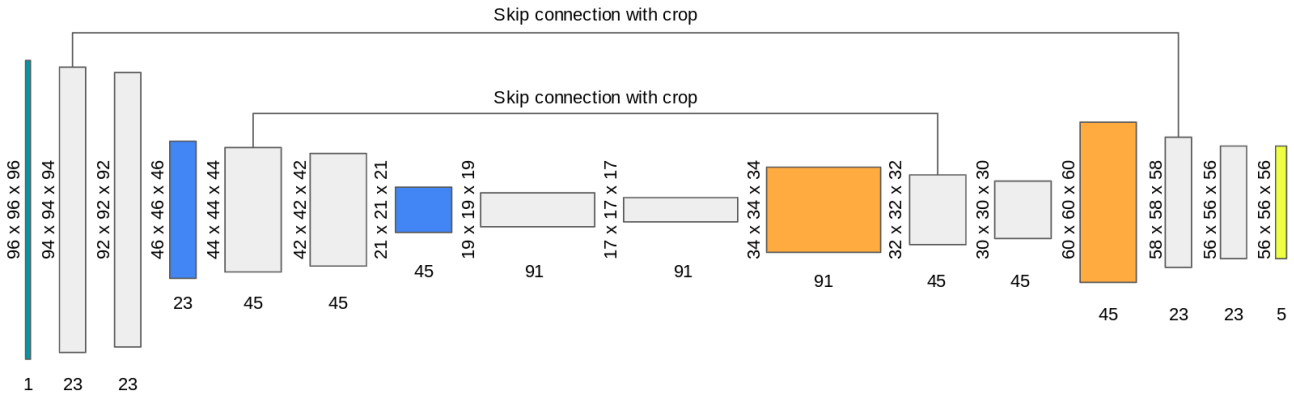

Figure 1. The 3D UNet-like network architecture: in teal is the input, in grey are 3D convolutions with stride 1 and no padding, in blue are downsampling convolutions with stride 2, in orange are plain upsampling layers with scale factor 2 and in yellow is the output. The numbers on the side of each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

The idea of transformation-consistency has been widely explored by recent research. In addition to Bortsova *et al.* and Li *et al.*,[2,3] on which our method is based, the Mean Teacher model[11] from classification has also been successfully adapted for image segmentation.[4,6] Liu *et al.*[5] 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.[12] Finally, Athiwaratkun *et al.*[13] 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.[2,3]

3. METHODOLOGY

3.1 Network Architecture

The network has a fully-convolutional[14] 3D Unet[15] architecture. 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[16] with skip connections. The convolutional layers use $3 \times 3 \times 3$ kernels, followed by batch normalisation[17] and leaky ReLU activation.[18] 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 3.2. The output dimensions are smaller than the input, so a segmentation is only produced for the center of the input image. Fig. 1 illustrates the architecture.

3.2 Training Procedure

This section describes the loss function, which follows the same structure as existing literature as well as Stochastic Weight Averaging (SWA), which we introduce for semi-supervised medical image segmentation.

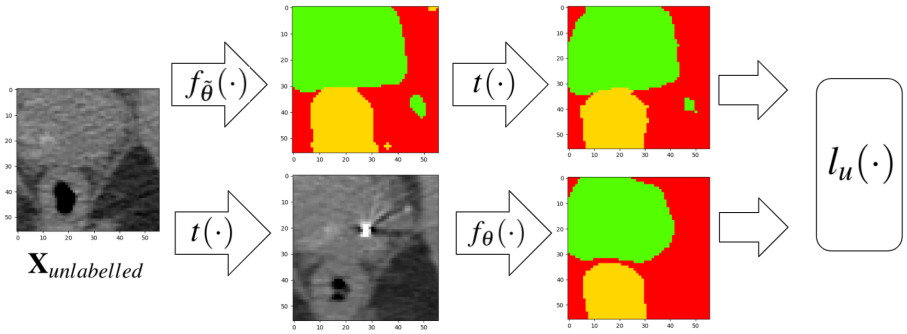

Figure 2. The unsupervised component, where the upper branch shows an unlabelled image being segmented then transformed, while the lower branch shows the same image being transformed then segmented. l_u is the unsupervised loss function.

3.2.1 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. 2, 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.*[3] but differs from Bortsova *et al.*,[2] 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.

3.2.2 Stochastic Weight Averaging

SWA was first introduced by Izmailov *et al.*,[8] 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,[19] which implies that the average of several solutions is likely to have a lower loss. Athiwaratkun *et al.*[13] 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.

3.2.3 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. 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.

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 $0.9 \times 0.9 \times 2 - 3mm$. 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.

4.2 Experimental Setup

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. Henceforth, experiments for the high data regime are labelled "105" and low data regime are labelled "20".

For each regime, results are presented for a fully-supervised (Base) and a semi-supervised network (SSL), plus SWA versions of both. Both the Base and SSL networks use the same UNet-like[15] architecture as described in Section 3.1. The difference is that λ in Eq. (1) is set to 0 for Base, 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 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. 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[20] with a constant maximum learning rate of 10^{-4} and the convolutional layers were initialised from a random normal distribution of $\mathcal{N}(0, 0.02^2)$.

All results are presented and evaluated in terms of mean surface distance (MSD) on the test set, while Dice and 95% Hausdorff Distance are enclosed in the supplement. We also compared the proposed approach to three state-of-the-art methods in abdominal CT radiotherapy: Cross-Stitch[21] is a deep learning approach that shares weights between a segmentation and registration CNN, Elastix[22] is a conventional iterative registration method and a Hybrid model[23] that feeds CNN segmentations of the bladder to an iterative approach as prior knowledge.

Table 1. Test set MSD (mm) values for the high and low labelled 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**

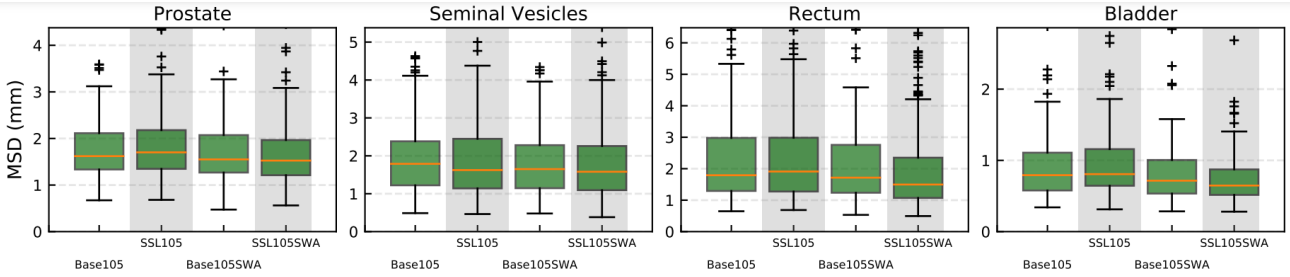

Figure 3. Distribution of test set MSD values for the high data regime.

4.3 Experimental Results

Table 1 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, 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. 3. 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 2 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.[21]

5. DISCUSSION AND CONCLUSION

This paper set out to investigate the effectiveness of SWA for transformation-consistent SSL in an image segmentation application. We also investigated the relative effective of the model in high and low labelled data regimes, as well as provided a comparison to previous state-of-the-art results on prostate CT datasets. The results for SSL without SWA, SSL20 and SSL105, are in line with existing literature,[2,3] 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, statistically

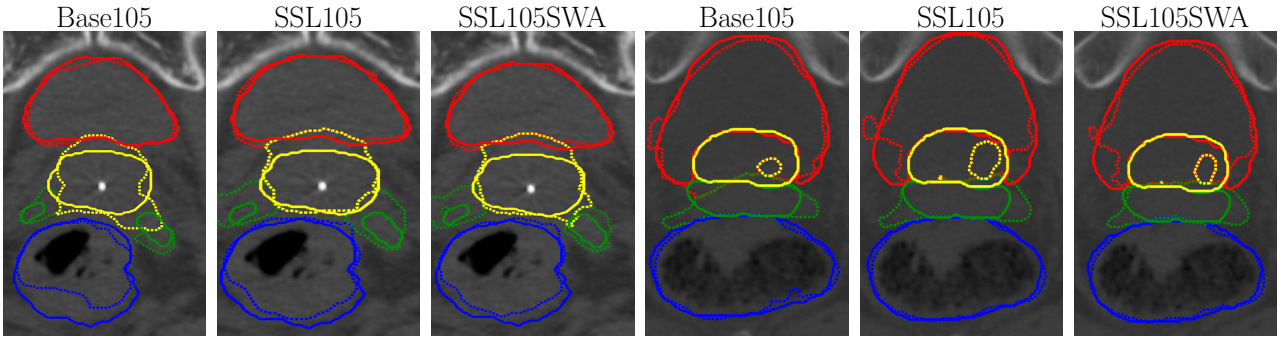

Figure 4. Segmentations from the high data regime for selected methods. The first three are from the first quartile in terms prostate MSD of the SSL105SWA network, while the last three are from the third quartile. The solid lines are groundtruth delineations by a radiation oncologist and the dotted lines are produced by the networks. Red, yellow, blue, and green represent the bladder, prostate, rectum, and seminal vesicles, respectively.

Table 2. 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*[21]	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[22]	$\mathbf{1.42} \pm 0.7$	**1.17**	2.07 ± 2.6	1.24	3.20 ± 1.6	3.07	5.30 ± 5.1	3.27
Hybrid[23]	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

significant improvement in both the high and low data regimes. Moreover, this is achievable without additional training cost, as discussed in Section 3.2.

To differentiate how much of the improvement is attributable to SSL and SWA respectively, supervised SWA models (Base20SWA and Base105SWA) were also provided for 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; a possible explanation could be that these two organs have larger degrees of daily variation, since their function is to fill and empty, which makes them particularly suitable for the transformation procedure embedded in our SSL method. Therefore, an avenue for further investigation could be the use of different transformation procedures for each organ. 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.

This paper compared the segmentation performance of a supervised baseline, a transformation-consistent semi-supervised model and SWA versions of both on a prostate CT dataset. The results showed that SSL + SWA also outperforms in the high labelled data regime, while existing research showed that SSL (without SWA) only outperforms in the low labelled data regime. 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.

Acknowledgements.

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.[24] and image samplers from NiftyNet[25] were used, for these we are also very grateful.

REFERENCES

[1] Figueroa, R., Zeng-Treitler, Q., Kandula, S., and Ngo, L., "Predicting sample size required for classification performance," *BMC Medical Informatics and Decision Making* **12**, 8 (2012).

[2] Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I., and de Bruijne, M., "Semi-supervised medical image segmentation via learning consistency under transformations," in [*Medical Image Computing and Computer Assisted Intervention*], 810–818, Springer International Publishing (2019).

[3] Li, X., Yu, L., Chen, H., Fu, C.-W., and Heng, P.-A., "Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model," in [*British Machine Vision Conference*], (2018).

[4] Li, X., Yu, L., Chen, H., Fu, C.-W., Xing, L., and Heng, P.-A., "Transformation consistent self-ensembling model for semi-supervised medical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems* **32**(2), 523–534 (2021).

[5] Liu, Q., Yu, L., Luo, L., Dou, Q., and Heng, P. A., "Semi-supervised medical image classification with relation-driven self-ensembling model," *IEEE Transactions on Medical Imaging* **39**(11), 3429–3440 (2020).

[6] Perone, C. S. and Cohen-Adad, J., "Deep semi-supervised segmentation with weight-averaged consistency targets," in [*International Workshop on Deep Learning in Medical Image Analysis*], *Lecture Notes in Computer Science* **11045**, 12–19 (2018).

[7] Sajjadi, M., Javanmardi, M., and Tasdizen, T., "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in [*NeurIPS*], (2016).

[8] Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G., "Averaging weights leads to wider optima and better generalization," in [*Conference on Uncertainty in Artificial Intelligence*], 876–885 (2018).

[9] Sonke, J.-J., Aznar, M., and Rasch, C., "Adaptive radiotherapy for anatomical changes," *Seminars in Radiation Oncology* **29**(3), 245–257 (2019).

[10] Cheplygina, V., de Bruijne, M., and Pluim, J. P., "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical Image Analysis* **54**, 280–296 (2019).

[11] Tarvainen, A. and Valpola, H., "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in [*Conference on Neural Information Processing System*], 1195–1204 (2017).

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

[13] Athiwaratkun, B., Finzi, M., Izmailov, P., and Wilson, A. G., "There are many consistent explanations of unlabeled data: Why you should average," in [*International Conference on Learning Representations*], (2019).

[14] Long, J., Shelhamer, E., and Darrell, T., "Fully convolutional networks for semantic segmentation," *IEEE Trans Pattern Anal Mach Intell* **39**(4), 640–651 (2017).

[15] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*Medical Image Computing and Computer-Assisted Intervention*], *Lecture Notes in Computer Science* **9351**, 234–241 (2015).

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

[17] Ioffe, S. and Szegedy, C., "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in [*Proceedings of Machine Learning Research*], **37**, 448–456 (2015).

[18] Maas, A. L., Hannun, A. Y., and Ng, A. Y., "Rectifier nonlinearities improve neural network acoustic models," in [*International Conference on Machine Learning*], (2013).

[19] Goodfellow, I. J., Vinyals, O., and Saxe, A. M., "Qualitatively characterizing neural network optimization problems," in [*International Conference on Learning Representations*], (2015).

[20] Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Han, J., "On the variance of the adaptive learning rate and beyond," in [*International Conference on Learning Representations*], (2020).

[21] Beljaards, L., Elmahdy, M. S., Verbeek, F., and Staring, M., "A cross-stitch architecture for joint registration and segmentation in adaptive radiotherapy," in [*Medical Imaging with Deep Learning*], *Proceedings of Machine Learning Research* **121**, 62 – 74 (2020).

[22] Qiao, Y., *Fast optimization methods for image registration in adaptive radiation therapy*, PhD thesis, Leiden University Medical Center (2017).

[23] Elmahdy, M. S., Jagt, T., Zinkstok, R. T., Qiao, Y., Shahzad, R., Sokooti, H., Yousefi, S., Incrocci, L., Marijnen, C., Hoogeman, M., and Staring, M., "Robust contour propagation using deep learning and image registration for online adaptive proton therapy of prostate cancer," *Medical physics* **46**(8), 3329–3343 (2019).

[24] Huang, C., Han, H., Yao, Q., Zhu, S., and Zhou, S. K., "3D U^2-net: A 3D universal U-net for multi-domain medical image segmentation," in [*Medical Image Computing and Computer Assisted Intervention*], *Lecture Notes in Computer Science* **11765**, 291–299 (2019).

[25] Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D. I., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y., Whyntie, T., Nachev, P., Modat, M., Barratt, D. C., Ourselin, S., Cardoso, M. J., and Vercauteren, T., "Niftynet: a deep-learning platform for medical imaging," *Computer Methods and Programs in Biomedicine* **158**, 113 – 122 (2018).