Can Generative Adversarial Networks help to overcome the limited data problem in segmentation?

Gerd Heilemann[1,2,*], **Mark Matthewman**[3], **Peter Kuess**[1,2], **Gregor Goldner**[1,2], **Joachim Widder**[1,2], **Dietmar Georg**[1,2], **Lukas Zimmermann**[1,4,5]

[1] Department of Radiation Oncology, Medical University of Vienna, Vienna, Austria
[2] Comprehensive Cancer Center, Medical University of Vienna, Vienna, Austria
[3] Technical University of Vienna, Vienna, Austria
[4] Competence Center for Preclinical Imaging and Biomedical Engineering, University of Applied Sciences Wiener Neustadt, Austria
[5] Faculty of Engineering, University of Applied Sciences Wiener Neustadt, Austria

Received 28 July 2021; accepted 23 November 2021

Abstract

Purpose: *For image translational tasks, the application of deep learning methods showed that Generative Adversarial Network (GAN) architectures outperform the traditional U-Net networks, when using the same training data size. This study investigates whether this performance boost can also be expected for segmentation tasks with small training dataset size.*

Materials/Methods: *Two models were trained on varying training dataset sizes ranging from 1—100 patients: a) U-Net and b) U-Net with patch discriminator (conditional GAN). The performance of both models to segment the male pelvis on CT-data was evaluated (Dice similarity coefficient, Hausdorff) with respect to training data size.*

Results: *No significant differences were observed between the U-Net and cGAN when the models were trained with the same training sizes up to 100 patients. The training dataset size had a significant impact on the models' performances, with vast improvements when increasing dataset sizes from 1 to 20 patients.*

Conclusion: *When introducing GANs for the segmentation task no significant performance boost was observed in our experiments, even in segmentation models developed on small datasets.*

Keywords: Automatic segmentation, Deep learning, Prostate cancer, Generative adversarial networks

1 Introduction

Segmentation of organs, parts of organs, and tumors on CT-data is an essential task in radiation oncology that is time consuming and prone to inaccuracies and inconsistencies [1]. The necessity of reliable and fast auto-segmentation has been widely discussed for years [2]. It has been shown that machine learning techniques can handle segmentation tasks of delineating organs-at-risk (OAR) in radiation oncology very well [3–6]. These provide huge potential in meeting the challenging demands of state-of-the-art radiation therapy concepts: One of the most intriguing trends in radiation oncology in recent years is towards adaptive strategies that aim at providing a more individualized therapy and improve the clinical outcome by accounting for anatomical changes between fractions or during treatment [7]. Adaptive radiotherapy is very resource demanding, with a great potential for automatization, especially in the repetitive task of delineating organs-at-risk (OAR).

The range of deep learning techniques used for segmentation in radiation oncology is wide, and some attempts have

[*] Corresponding author: Gerd Heilemann, Department of Radiation Oncology, Medical University of Vienna, Vienna, Austria.
E-mail: gerd.heilemann@meduniwien.ac.at (G. Heilemann).

https://doi.org/10.1016/j.zemedi.2021.11.006
www.elsevier.com/locate/zemedi

been made to provide some sort of comparability among the different models [5,8–10]. However, the most common models, also the more recent ones, such as the very successful nnU-Net by Isensee *et al* [11], are U-Net-based architectures, introduced by Ronneberger *et al* in 2015 [12], which have been applied for many segmentation tasks in radiation oncology [13–18].

Training data for medical images is often limited by the amount of available high quality labelled clinical data which has a direct impact on the performance of the learning algorithm [19]. Therefore, using a learning algorithm which can maintain the accuracy with reduced training data would be an asset for clinical application.

Particularly, in the domain of translational research, Generative Adversarial Networks (GANs) [20] are combined with classical architectures (e.g. U-Net) to improve performance either by relying on the enhanced metric computation or by generalizing with small training datasets [15,21,22]. Especially for segmentation the concept of conditional GANs (cGAN) is important as a direct correlation between the labels and the image can increase training performance as it was demonstrated by Isola *et al* [23]. Thus, it would be expected that a cGAN achieves the required performance of a U-Net architecture with a lower size of training data. The concept of requiring less data for training deep learning models and still achieving acceptable performance would help to focus on collecting high quality datasets instead of big datasets including poorly annotated data.

In this study, a U-Net and a cGAN based model were investigated for male pelvic OAR segmentation to assess the potential benefit of cGANs for small training set sizes. The influence of an additional discriminative network was analyzed and compared to the standard U-Net architecture. The focus was to evaluate the performance of different training dataset sizes.

2 Methods

2.1 Data

In total, data from 308 prostate cancer patients were included in this study. All patients were treated at the Department of Radiation Oncology at the Medical University of Vienna, between 2016 and 2018. Ethical approval was granted by the institutional ethics committee (1255/2021). A set of 100 patients was randomly selected for training. To investigate the impact of different training dataset sizes, this training set was split into seven subsets of different size (see paragraph 2.3). The validation dataset consisted of 29 patients to validate the results of the different models. The overall best models were tested on a cohort of 179 patients.

All patients received a CT scan on a Siemens Somatom (Siemens Healthineers, Erlangen, Germany) prior treatment with a resolution of 512 × 512 pixel, a median pixel spacing of 0.93 mm (range: 0.70–1.27 mm), a slice thickness of 2 mm and a pelvic protocol. The delineation of the target and OARs (i.e. bladder, rectum, and femoral heads) was performed by a highly experienced radiation oncologist (G.G.) with more than 20 years of experience in prostate radiotherapy.

2.2 Neural network architecture and training

Two network architectures were analyzed (see Figure 1): a) U-Net [12,23] and b) U-Net combined with a patch discriminator (cGAN) [23,24]. The model was trained slice-by-slice to segment the following structures: bladder, rectum, left and right femoral heads, external body contour and exterior (i.e. air). In both cases the generator model was the same, including a combination of convolution, normalization (instance normalization) and activation blocks (ReLU or LeakyReLU with slope 0.2) for nine down-convolution paths. In the bottleneck the image was compressed to a $1 \times 1 \times 512$ vector. A kernel size of 4×4 was applied with 64 features in the first layer. The features were doubled when the image dimension was reduced. The feature map number was kept constant when 512 features were reached. Input consisted of single transversal slices with an image size of 512×512 pixels. The model output size was $512 \times 512 \times 6$ pixels (with each channel corresponding to one structure). A Softmax function was applied to give the final probability for each class. The discriminator for the cGAN model was a patch discriminator which reduced the image size to a 60×60 feature map. This feature map of the discriminator was compared with a tensor of the same size filled with 1 if the input image was a real image pair and with zeros if it was the prediction. This was used as input for classification with the CT image and the masks (predicted or ground truth) as concatenated input. The final receptive field was 70×70 pixels, meaning that one pixel of the feature map sees a 70×70 field. Similar to the generator convolution, normalization, and activation blocks were used.

Models were trained with a batch size of three images for 100 epochs, where the learning rate was constant for the first 50 epochs and then linearly decreased over the last epochs (down to zero after 100 epochs). Adam was used for optimization, with a learning rate of $2 \times 10{-}4$, and fixed β values of 0.9 and 0.999 [25]. All weights were initialized using the Kaiming method [26]. Batch size and number of epochs were fixed after an initial pilot experiment.

To increase the sample size, random horizontal flips were included. If the image was flipped the right and left femoral head labels were swapped.

All models were implemented using the PyTorch library (version 1.0). Training on 26 cases took up to 14 hours on a NVIDIA Geforce 1080Ti with 11 GB memory.

A comprehensive hyper-parameter search was conducted, including different loss metrics and weighting between those loss metrics to find the optimal settings per dataset size. The data was reviewed to identify the pareto optimal surface.

The Dice-similarity-coefficient (DSC) loss and the cross-entropy (CE) loss were implemented and used as loss metrics.

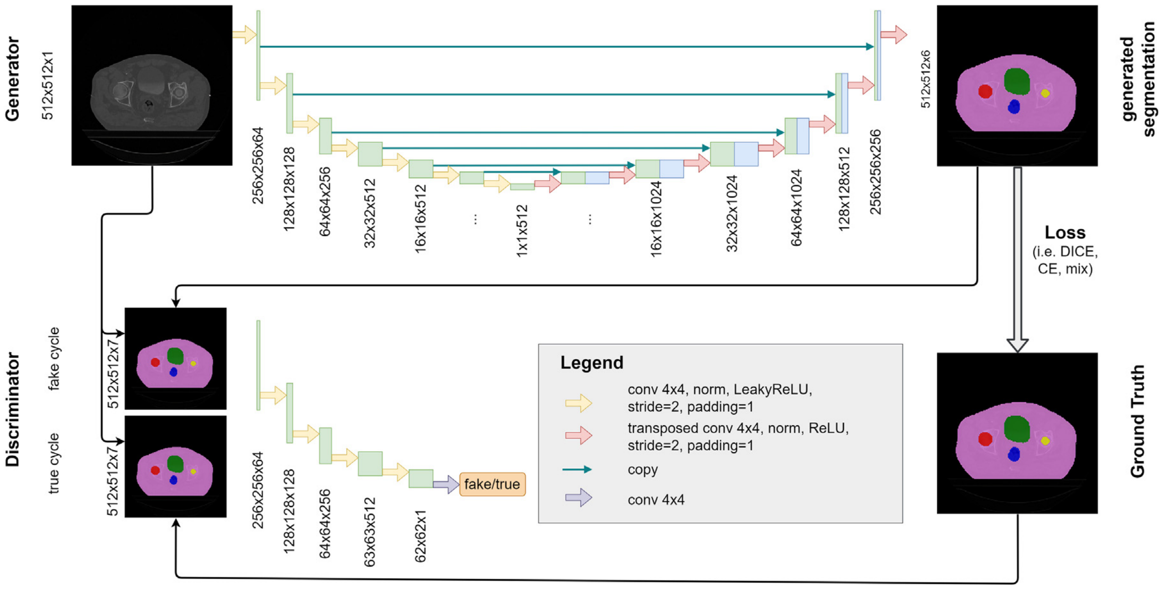

Figure 1. Network architectures of the U-Net (top only: Generator) and the GAN (top and bottom: generator + discriminator).

Additionally, different weightings between these loss metrices were used to identify their influence on the validation performance. The regularization rate λ was set to 0, 0.25, 0.5, 0.75, and 1 where 0 corresponded to pure DSC and 1 to pure CE loss. Individual structure classes were weighted differently for the CE loss as a class imbalance was given for labels like femoral bones and body. Initially, the weightings for all classes were 1, which represents the original cross-entropy loss. Secondly, a weighting of 2.5 was introduced for the – compared to body and air – relatively small structures (i.e. femoral heads, bladder, and rectum). In a third step, a dynamic weighting was implemented, which included the information on how well the auto-segmentation is performing during the training. This can be represented as:

$$L_{combined} = (1 - \lambda) \cdot (1 - DSC) + \lambda \cdot CE \qquad (1)$$

where DSC and CE losses are given as:

$$DSC = \frac{1}{6}\sum_{i}^{n=6} \frac{2TP[i]}{2TP[i] + FP[i] + FN[i]} \qquad (2)$$

and

$$CE\,(x, i) = w_i \left(-x_i + \log\left(\sum_j e^{x_j} \right) \right) \qquad (3)$$

with w_i being the weight defined for each organ, n the number of classes (n = 6), $TPFP$ and FN the rate of true/false positive and false negative classifications. The dynamic weighting of CE was implemented as in [27] to weight single class values higher in the presence of a larger mismatch:

$$w_i = \frac{N_{total}}{TP[i]} \cdot 100 \qquad (4)$$

where N_{total} is the number of total voxels and $TP[i]$ the true positives within each class i.

Table 1
Data split for the different experiments showing the number of patients for each set.

Training	Validation (fixed)	Testing (fixed)
1) 1	29	179
2) 6		
3) 11		
4) 16		
5) 21		
6) 26		
7) 100		

2.3 Training set size

The training dataset was divided into training set sizes ranging between 1 to 26 in steps of 5 patients (see Table 1). For each of these sets the hyperparameter search was repeated to determine the optimal loss setting. Lastly, a final model was trained on 100 patients, to analyze the performance of both networks on a reasonably sized training set.

2.4 Evaluation metrics and statistical analysis

The testing was performed with the model state after 100 epochs. The models were selected by the highest DSC, which is the most widely accepted metric among deep learning-based auto-segmentation studies [28]. The model performance was validated for each organ separately using the DSC metric, recall, precision, MSE, and 95% Hausdorff distance of the segmentation masks.

The influence of the training data size was investigated by using the best performing models for all organs and all dataset sizes to determine the optimal settings for the weighting between DSC and CE. The metric results of the validation dataset were plotted to analyze the impact of the additional GAN metric and the data size.

Friedman tests (and post hoc Nemenyi tests) were performed to test for significant differences between the different settings and models. All statistical tests were done in Python using the SciPy and scikit-learn libraries.

3 Results

3.1 Loss metrics

The implementation of the different losses showed a trend that for larger training datasets CE weighting produced better results, whereas smaller training dataset sizes benefitted from dynamic weighting. However, the search was not conclusive leading to a non-significant difference between the different settings which was tested with the Friedman test ($p > 0.05$). In general, a CE loss with 0.25 weighted DSC metric resulted in better performance.

3.2 Impact of training dataset size

In Figure 2 the DSC metrics are visualized showing a sharp increase of the performance between 1 and 6 training patients with a small increase after gradually including more patients. Most pronounced for the rectum, where the performance from 1 to 6 patients increased by a DSC score of 0.6. The difference between cGAN and U-Net metrics among each OAR was small (<2%) for all training dataset sizes. The interquartile range of both models overlapped for all dataset sizes. Stepwise increasing the training dataset size yielded significant differences (Nemenyi post hoc test) for some, but not all increments. However, the effect between training size increases was always larger than changing the network (see below).

3.3 U-Net vs. cGAN

No significant differences (Friedman $p \gg 0.05$) were observed between the cGAN and the U-Net within one dataset size (see Table 2).

Figure 3 shows the correlation between ground truth and predicted volume of the model trained with 100 patients. The volumes of bladder and rectum tend to be underrepresented. The largest differences between cGAN and U-Net were found in the rectum where R2 is 0.79 for U-Net and 0.83 for cGAN. These differences almost vanished for the other structures, which yielded almost identical R2 values between the models.

4 Discussion

Reliable and accurate segmentation of OARs is important for radiation therapy treatment planning and automated segmentation methods become key when aiming for online adaptive treatment strategies [2]. Over recent years, the introduction of deep learning methods for segmentation tasks in radiation therapy showed very promising results, matching or even outperforming state-of-the-art deformable model based or atlas based techniques [3,13,28].

Results of our model were comparable to recent studies [13,14,17,18], but showed no benefit when including an additional discriminative model into the training process. The hypothesis that a cGAN architecture provides better results for small dataset sizes in comparison to a more general and simpler U-Net architectures could not be confirmed. No significant differences between both architectures for structure metrics (e.g. DSC, sensitivity, Hausdorff distance) were found.

Dong *et al* claimed that the training results stabilized after 20 patients for a very similar cGAN architecture as used in this study [15]. But they did not compare to any other network. Our study confirms a stabilization around 20 patients for both cGAN and U-Net; however, continuous improvement can be observed if significantly larger datasets are included in the training. In a very recent review by Vandewinckele *et al* [29] they found that most state-of-the-art CNN-based contouring models were trained on 100 or more patients, but, they state that some studies achieved reasonably good results with as little as 50 patients. The performance analysis over the different dataset sizes demonstrates the importance of large training datasets and the importance of the loss metric dependent on the training data size. However, the highest metric improvements can be observed in the range of 1–21 patients in this study.

Some of the most recent studies have implemented GAN structures, consisting of a U-Net-like generator and a discriminator (e.g. fully convolutional networks) for the segmentation of male pelvis [30]. Particularly Sultana *et al* [17] showed results well in line with our model. They used a similar U-Net and GAN structure for multiclass segmentation and trained on the same number of patients as our final model. While our study cannot confirm the performance boost with respect to DSC, we observed a similar tendency in the Hausdorff distance. Similar cGAN architectures have been applied for multi-organ segmentation in other regions, e.g. in the thorax [15]. Several studies showed that the U-Net was very versatile, with an overall accurate segmentation performance for a

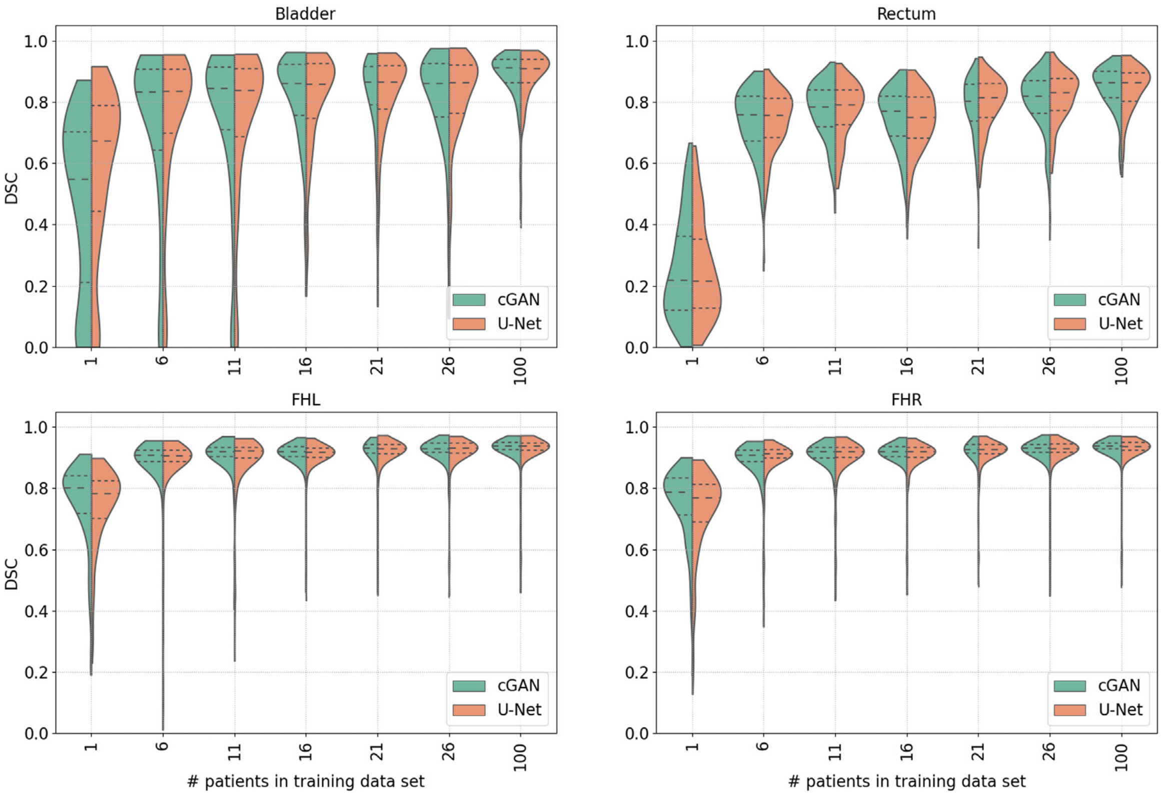

Figure 2. Comparison of the performance of the two networks cGAN (left, green) and U-Net (right, red) with respect to the different training data sizes. The DSC results are depicted for the bladder, rectum, left femoral head (FHL) and right femoral head (FHR).

Table 2
Comparison of the performance of the U-Net and cGAN for different metrics. The values are from the model trained on the largest training dataset used in this study (100 patients). Statistical analyses showed no significant differences between U-Net and cGAN.

Organ	Metric	DSC	Precision	Sensitivity	Hausdorff distance
Rectum	U-Net	0.84 ± 0.08	0.85 ± 0.11	0.84 ± 0.09	7.77 ± 6.62
	GAN	0.84 ± 0.08	0.85 ± 0.11	0.85 ± 0.09	6.79 ± 5.08
Bladder	U-Net	0.88 ± 0.09	0.90 ± 0.10	0.88 ± 0.11	6.90 ± 13.18
	GAN	0.89 ± 0.08	0.90 ± 0.09	0.88 ± 0.11	6.02 ± 6.95
Femoral head left	U-Net	0.93 ± 0.06	0.93 ± 0.08	0.93 ± 0.05	4.06 ± 8.41
	GAN	0.93 ± 0.06	0.93 ± 0.08	0.93 ± 0.05	3.53 ± 4.35
Femoral head right	U-Net	0.93 ± 0.06	0.93 ± 0.08	0.93 ± 0.05	3.66 ± 4.48
	GAN	0.93 ± 0.06	0.93 ± 0.08	0.93 ± 0.05	3.59 ± 4.44

wide range of anatomical regions [16,31]. While our study is limited to the male pelvic, we expect the results to be equally extrapolatable.

GANs have provided impressive results in the field of computer vision [23,32]. Particularly improvements in the data efficient training played a major role recently, introducing augmentation methods which can be applied directly on the discriminator without leaking these transforms towards the generator [33,34]. Especially for medicine, these techniques could be interesting as many color and spatial transformations cannot be used because the image context could be disrupted. Whether the advantages of GANs exist in the same order of magnitude for medical images – and radiation therapy in particular – is still to be investigated [21,35]. However, the availability of training data for medical images is orders of magnitude lower than for other computer vision tasks, providing opportunities for these data-efficient training techniques. Recent studies have highlighted different strategies to overcome this limited data problem, e.g. by more sophisticated data augmentation [36] or the generation of artificial training data [37–39].

In general, high cross-entropy weighting produced more favorable results with respect to the investigated evaluation metrics. The overestimation, false positive classification of

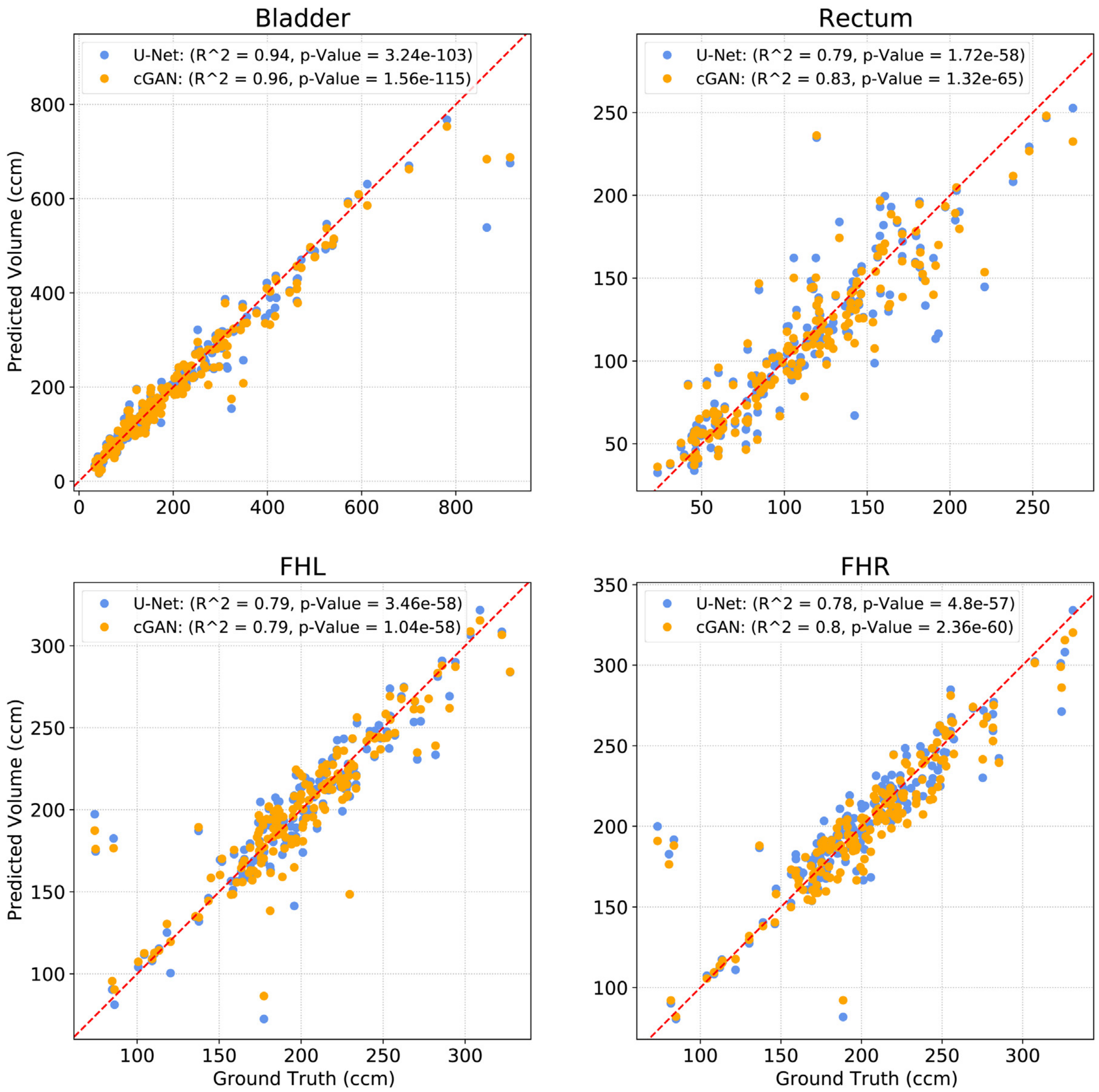

Figure 3. Predicted vs. delineated structure volume for bladder, rectum, left femoral head (FHL) and right femoral head (FHR).

pixels, is a consequence of the DSC in the regularization term of the loss.

In this study, no multi-reader approach was used, but all delineation was performed by the same oncologist. While this is sub-optimal when aiming for a model with peak performance, it was beneficial in this study, as the ground truth is of very consistent quality and thus the comparison between both approaches can be focused on the network architecture.

The systematic underestimation of rectum and bladder volumes might be caused by difficulties of the network in dealing with the inconsistent delineation start and end points in cranial and caudal direction of the input data. For the bladder this is amplified by the low cranio-caudal resolution and the slice based 2D training of the networks. Additionally, different sample-mining (e.g. ratio-based weighting [40]) approaches might help overcome the under-representation problem of relatively small volumes.

The results of this study did not show noticeable improvements of GAN training with respect to the metric results of segmentation maps independent of the training sample size in comparison to a standard U-Net architecture and might not justify the implementation of a discriminative network. The memory used for the discriminator should be more efficiently allocated to increase the model size and push the model to more accurate results. Instead, more sophisticated methods to overcome the limited data problem are necessary [33] (e.g. improved augmentation methods, application of pretrained models), which in turn are very costly in terms of training.

5 Conclusion

This study investigated if automated segmentation tasks in the male pelvic benefit from cGAN architectures in comparison to a U-Net architecture, focusing on the size of the training dataset. The results do not yield arguments to apply a cGAN for these tasks or to claim that U-Net based results might improve when changing the architecture to cGAN.

Appendix A Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.zemedi.2021.11.006.

Literature

[1] Nelms BE, Tomé WA, Robinson G, Wheeler J. Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer. Int J Radiat Oncol Biol Phys 2012;82:368–78, http://dx.doi.org/10.1016/j.ijrobp.2010.10.019.

[2] Sharp G, Fritscher KD, Pekar V, Peroni M, Shusharina N, Veeraraghavan H, et al. Vision 20/20: perspectives on automated image segmentation for radiotherapy. Med Phys 2014;41:1–13, http://dx.doi.org/10.1118/1.4871620.

[3] Seo H, Badiei Khuzani M, Vasudevan V, Huang C, Ren H, Xiao R, et al. Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications. Med Phys 2020;47:e148–67, http://dx.doi.org/10.1002/mp.13649.

[4] Elguindi S, Zelefsky MJ, Jiang J, Veeraraghavan H, Deasy JO, Hunt MA, et al. Deep learning-based auto-segmentation of targets and organs-at-risk for magnetic resonance imaging only planning of prostate radiotherapy. Phys Imaging Radiat Oncol 2019;12:80–6, http://dx.doi.org/10.1016/j.phro.2019.11.006.

[5] Raudaschl PF, Zaffino P, Sharp GC, Spadea MF, Chen A, Dawant BM, et al. Evaluation of segmentation methods on head and neck CT: auto-segmentation challenge 2015. Med Phys 2017;44:2020–36, http://dx.doi.org/10.1002/mp.12197.

[6] van der Veen J, Willems S, Deschuymer S, Robben D, Crijns W, Maes F, et al. Benefits of deep learning for delineation of organs at risk in head and neck cancer. Radiother Oncol 2019;138:68–74, http://dx.doi.org/10.1016/j.radonc.2019.05.010.

[7] McPartlin AJ, Li XA, Kershaw LE, Heide U, Kerkmeijer L, Lawton C, et al. MRI-guided prostate adaptive radiotherapy – a systematic reviewMRI-linac and prostate motion review. Radiother Oncol 2016;119:371–80, http://dx.doi.org/10.1016/j.radonc.2016.04.014.

[8] Qin W, Wu J, Han F, Yuan Y, Zhao W, Ibragimov B, et al. Superpixel-based and boundary-sensitive convolutional neural network for automated liver segmentation. Phys Med Biol 2018:63, http://dx.doi.org/10.1088/1361-6560/aabd19.

[9] Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 2015;34:1993–2024, http://dx.doi.org/10.1109/TMI. 2014.2377694.

[10] Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Z Med Phys 2019;29:102–27, http://dx.doi.org/10.1016/j.zemedi.2018.11.002.

[11] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18:203–11, http://dx.doi.org/10.1038/s41592-020-01008-z.

[12] Ronneberger O, Fischer P, Brox T. U-net:Convolutional networks for biomedical image segmentation. Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2015;9351:234–41, http://dx.doi.org/10.1007/978-3-319-24574-4_28.

[13] Balagopal A, Kazemifar S, Nguyen D, Lin MH, Hannan R, Owrangi A, et al. Fully automated organ segmentation in male pelvic CT images. ArXiv 2018.

[14] Wang S, He K, Nie D, Zhou S, Gao Y, Shen D. CT male pelvic organ segmentation using fully convolutional networks with boundary sensitive representation. Med Image Anal 2019;54:168–78, http://dx.doi.org/10.1016/j.media.2019.03.003.

[15] Dong X, Lei Y, Wang T, Thomas M, Tang L, Curran WJ, et al. Automatic multiorgan segmentation in thorax CT images using U-net-GAN. Med Phys 2019;46:2157–68, http://dx.doi.org/10.1002/mp.13458.

[16] Nikolov S, Blackwell S, Mendes R, De Fauw J, Meyer C, Hughes C, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy; 2018. p. 1–31.

[17] Sultana S, Robinson A, Song DY, Lee J. CNN-based hierarchical coarse-to-fine segmentation of pelvic CT images for prostate cancer radiotherapy. In: Fei B, Linte CA, editors. Med. Imaging 2020 Image-Guided Proced. Robot. Interv. Model. SPIE; 2020. p. 53, http://dx.doi.org/10.1117/12.2549979.

[18] Kazemifar S, Balagopal A, Nguyen D, McGuire S, Hannan R, Jiang S, et al. Segmentation of the prostate and organs at risk in male pelvic CT images using deep learning. Biomed Phys Eng Expr 2018:4, http://dx.doi.org/10.1088/2057-1976/aad100.

[19] Shen C, Nguyen D, Zhou Z, Jiang SB, Dong B, Jia X. An introduction to deep learning in medical physics: advantages, potential, and challenges. Phys Med Biol 2020:65, http://dx.doi.org/10.1088/1361-6560/ab6f51.

[20] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Adv Neural Inf Process Syst 2014;3:2672–80.

[21] Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: a review. Med Image Anal 2019;58:1–20, http://dx.doi.org/10.1016/j.media.2019.101552.

[22] Tajbakhsh N, Jeyaseelan L, Li Q, Chiang JN, Wu Z, Ding X. Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. Med Image Anal 2020;63:101693, http://dx.doi.org/10.1016/j.media.2020.101693.

[23] Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017. 2017. p. 5967–76, http://dx.doi.org/10.1109/CVPR. 2017.632.

[24] Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proc IEEE Int Conf Comput Vis 2017;2017(October):2242–51, http://dx.doi.org/10.1109/ICCV. 2017.244.

[25] Kingma DP, Ba JL. Adam: a method for stochastic optimization. In: 3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc. 2015. p. 1–15.

[26] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. Proc IEEE Int Conf Comput Vis 2015;2015(Inter):1026–34, http://dx.doi.org/10.1109/ICCV. 2015.123.

[27] Lu S, Gao F, Piao C, Ma Y. Dynamic weighted cross entropy for semantic segmentation with extremely imbalanced data. In: 2019 Int Conf. Artif. Intell. Adv. Manuf., IEEE. 2019. p. 230–3, http://dx.doi.org/10.1109/AIAM48774.2019.00053.

[28] Vrtovec T, Močnik D, Strojan P, Pernuš F, Ibragimov B. Auto-segmentation of organs at risk for head and neck radiotherapy planning: from atlas-based to deep learning methods. Med Phys 2020:47, http://dx.doi.org/10.1002/mp.14320.

[29] Vandewinckele L, Claessens M, Dinkla A, Brouwer C, Crijns W, Verellen D, et al. Overview of artificial intelligence-based applications in radiotherapy: recommendations for implementation and quality assurance. Radiother Oncol 2020;153:55–66, http://dx.doi.org/10.1016/j.radonc.2020.09.008.

[30] Fu Y, Lei Y, Wang T, Tian S, Patel P, Jani AB, et al. Pelvic multi-organ segmentation on cone-beam CT for prostate adaptive radiotherapy. Med Phys 2020;47:3415–22, http://dx.doi.org/10.1002/mp.14196.

[31] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88, http://dx.doi.org/10.1016/j.media.2017.07.005.

[32] Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of stylegan. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2020:8107–16, http://dx.doi.org/10.1109/CVPR42600.2020.00813.

[33] Karras T, Aittala M, Hellsten J, Laine S, Lehtinen J, Aila T. Training generative adversarial networks with limited data. ArXiv 2020. http://arxiv.org/abs/2006.06676.

[34] Zhao S, Liu Z, Lin J, Zhu J-Y, Han S. Differentiable augmentation for data-efficient GAN training. ArXiv 2020. http://arxiv.org/abs/2006.10738.

[35] Lan L, You L, Zhang Z, Fan Z, Zhao W, Zeng N, et al. Generative adversarial networks and its applications in biomedical informatics. Front Public Heal 2020;8:1–14, http://dx.doi.org/10.3389/fpubh.2020.00164.

[36] Peng Z, Zhou J, Fang X, Yan P, Shan H, Wang G, et al. Data augmentation for training deep neural networks auto-segmentation. Radiat Oncol 2021:151–64, http://dx.doi.org/10.1201/9780429323782-13.

[37] Russ T, Goerttler S, Schnurr AK, Bauer DF, Hatamikia S, Schad LR, et al. Synthesis of CT images from digital body phantoms using CycleGAN. Int J Comput Assist Radiol Surg 2019;14:1741–50, http://dx.doi.org/10.1007/s11548-019-02042-9.

[38] Bauer DF, Russ T, Waldkirch BI, Tönnes C, Segars WP, Schad LR, et al. Generation of annotated multimodal ground truth datasets for abdominal medical image registration. Int J Comput Assist Radiol Surg 2021;16:1277–85, http://dx.doi.org/10.1007/s11548-021-02372-7.

[39] Fetty L, Bylund M, Kuess P, Heilemann G, Nyholm T, Georg D, et al. Latent space manipulation for high-resolution medical image synthesis via the StyleGAN. Z Med Phys 2020;30:305–14, http://dx.doi.org/10.1016/j.zemedi.2020.05.001.

[40] Golla A-K, Bauer DF, Schmidt R, Russ T, Norenberg D, Chung K, et al. Convolutional neural network ensemble segmentation with ratio-based sampling for the arteries and veins in abdominal CT scans. IEEE Trans Biomed Eng 2021;68:1518–26, http://dx.doi.org/10.1109/TBME. 2020.3042640.