V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation

Fausto Milletari
Technical University of Munich
Boltzmannstr. 3, 85748 Munich
fausto.milletari@tum.de

Nassir Navab
Technical University of Munich
Boltzmannstr. 3, 85748 Munich
navab@cs.tum.edu

Seyed-Ahmad Ahmadi
Ludwig-Maximilians-Universität München
Marchioninistrasse 15, 81377 Munich
ahmad.ahmadi@med.uni-muenchen.de

Abstract

Convolutional Neural Networks (CNNs) have been recently employed to solve problems from both the computer vision and medical image analysis fields. Despite their popularity, most approaches are only able to process 2D images while most medical data used in clinical practice consists of 3D volumes. In this work we propose an approach to 3D image segmentation based on a volumetric, fully convolutional, neural network. Our CNN is trained end-to-end on MRI volumes depicting prostate, and learns to predict segmentation for the whole volume at once. We introduce a novel objective function, that we optimise during training, based on Dice coefficient. In this way we can deal with situations where there is a strong imbalance between the number of foreground and background voxels. To cope with the limited number of annotated volumes available for training, we augment the data applying random non-linear transformations and histogram matching. We show in our experimental evaluation that our approach achieves good performances on challenging test data while requiring only a fraction of the processing time needed by other previous methods.

1. Introduction

Recent research in computer vision and pattern recognition has highlighted the capabilities of Convolutional Neural Networks (CNNs), achieving state-of-the-art performances on challenging tasks such as classification, segmentation and object detection. This success has been attributed to the ability of CNNs to learn a hierarchical representation of raw input data, without relying on handcrafted features. As the inputs are processed through the network layers, the level of abstraction of the resulting features increases. Shallower layers grasp local information while deeper layers use filters with much broader receptive fields, able to capture more global information [23].

Segmentation is a highly relevant task in medical image analysis. Automatic delineation of organs and structures of interest is often necessary to perform tasks such as visual augmentation [13], computer assisted diagnosis [15], interventions [24] and extraction of quantitative indices from images [1]. However, compared to 2D images mostly used in computer vision, diagnostic and interventional images data in the medical field are often volumetric. This creates a need for algorithms performing segmentations in 3D, by taking the whole volume content into account at once.

In this work, we aim to segment prostate MRI volumes. This task is clinically relevant both during diagnosis, e.g. due to volume assessment [16], and during treatment planning, e.g. by accurate boundary estimation [7, 24]. Prostate segmentation from MRI can be challenging due to large appearance variation across different scans, e.g. in terms of deformations or changes of the intensity distribution. Moreover, MRI volumes are often affected by artefacts and distortions due to field inhomogeneity.

In this work we present a novel 3D segmentation approach that leverages the power of a fully convolutional neural network, trained end-to-end, for processing of volumetric medical images such as MRI. Compared to other recent approaches, our contributions are three-fold. First, instead of processing the input volumes in a 2D slice-by-slice fashion, we propose to directly use 3D convolutions. Second, we propose to maximize a novel objective function designed specifically for medical image segmentation, which

978-1-5090-5407-7/16 $31.00 © 2016 IEEE
DOI 10.1109/3DV.2016.79

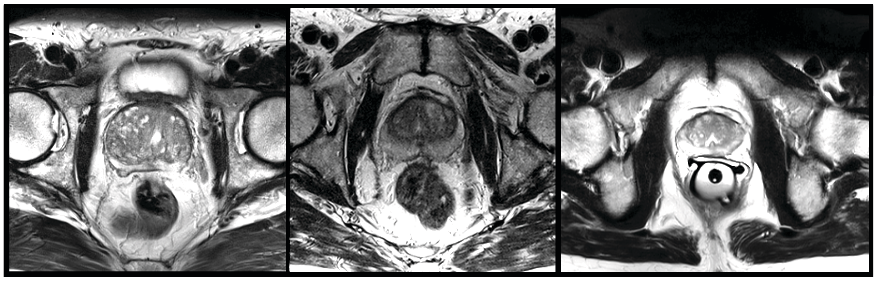

Figure 1. Slices from MRI volumes depicting prostate. This data is part of the PROMISE2012 challenge dataset [10].

is based on the Dice overlap coefficient. Third, we integrate recent insights into improved training convergence from literature by formulating each convolutional stage such that it learns a residual function. We empirically observed that this mechanism ensures also our novel architecture to converge in a fraction of the time required by a similar network that does not learn residuals, when applied to the segmentation task on the Promise 2012 dataset. We demonstrate fast and accurate results on prostate MRI test volumes and we provide direct comparison with other methods which were evaluated on the same test data[1].

2. Related Work

CNNs have been recently used for medical image segmentation. Early approaches obtain anatomy delineation in images or volumes by performing patch-wise image classification. Such segmentations are obtained by only considering local context and therefore are prone to failure, especially in challenging modalities such as ultrasound, where a high number of mis-classified voxel are to be expected. Post-processing approaches such as connected components analysis normally yield no improvement and therefore, more recent works, propose to use the network predictions in combination with Markov random fields [9], voting strategies [12] or more traditional approaches such as level-sets [2]. Patch-wise approaches also suffer from efficiency issues. When densely extracted patches are processed in a CNN, a high number of computations is redundant and therefore the total algorithm runtime is high. In this case, more efficient computational schemes can be adopted. CNNs employing fully connected layers can be, in some situations and with a few restrictions, converted to fully convolutional ones as shown in [18].

Fully convolutional networks trained end-to-end were previously applied to 2D images both in computer vision [14, 11] and microscopy image analysis [17]. These models, which served as an inspiration for our work, employed fully convolutional network architectures and were trained to predict a segmentation mask, delineating the structures of interest, for the whole image. In [14], a pre-trained VGG network architecture [19] was used in conjunction with its mirrored, de-convolutional equivalent to segment RGB images, by leveraging the descriptive power of the features extracted by the innermost layer. In [11], three fully convolutional deep neural networks, pre-trained on a classification task, were refined to produce segmentations while in [17], a brand new CNN model, especially tailored to tackle biomedical image analysis problems in 2D, was proposed. More recently, this approach was extended to 3D and applied to segmentation of volumetric data acquired from a confocal microscope [3]. The method was trained using partially annotated data, by optimization of a weighted multinomial logistic loss layer. Compared to this, we introduce a novel loss layer specifically designed for segmentation tasks, which is based on Dice coefficient, one of the most common measures of region overlap in medical image analysis [4]. Our experiments show that a direct optimization of this objective overlap measure during training yields better segmentation accuracy than the commonly used multinomial loss function, as used e.g. in [3].

3. Method

In Figure 2 we provide a schematic representation of our convolutional neural network. We perform convolutions aiming both at extracting features from the data and, at the end of each stage, reducing its resolution by using appropriate stride. The left part of the network consists of a compression path, while the right part decompresses the signal until its original size is reached. Convolutions are all applied with appropriate padding.

The left side of the network is divided into different stages that operate at different resolutions. Each stage comprises one to three convolutional layers. Similarly to the approach presented in [5], we formulate each stage such that it learns a residual function: the input of each stage is (a) used in the convolutional layers and processed through the non-linearities and (b) added to the output of the last convolutional layer of that stage in order to enable learning a residual function. As confirmed by our empirical observations, this architecture ensures convergence in a fraction of

[1] Detailed results available on http://promise12.grand-challenge.org/results/

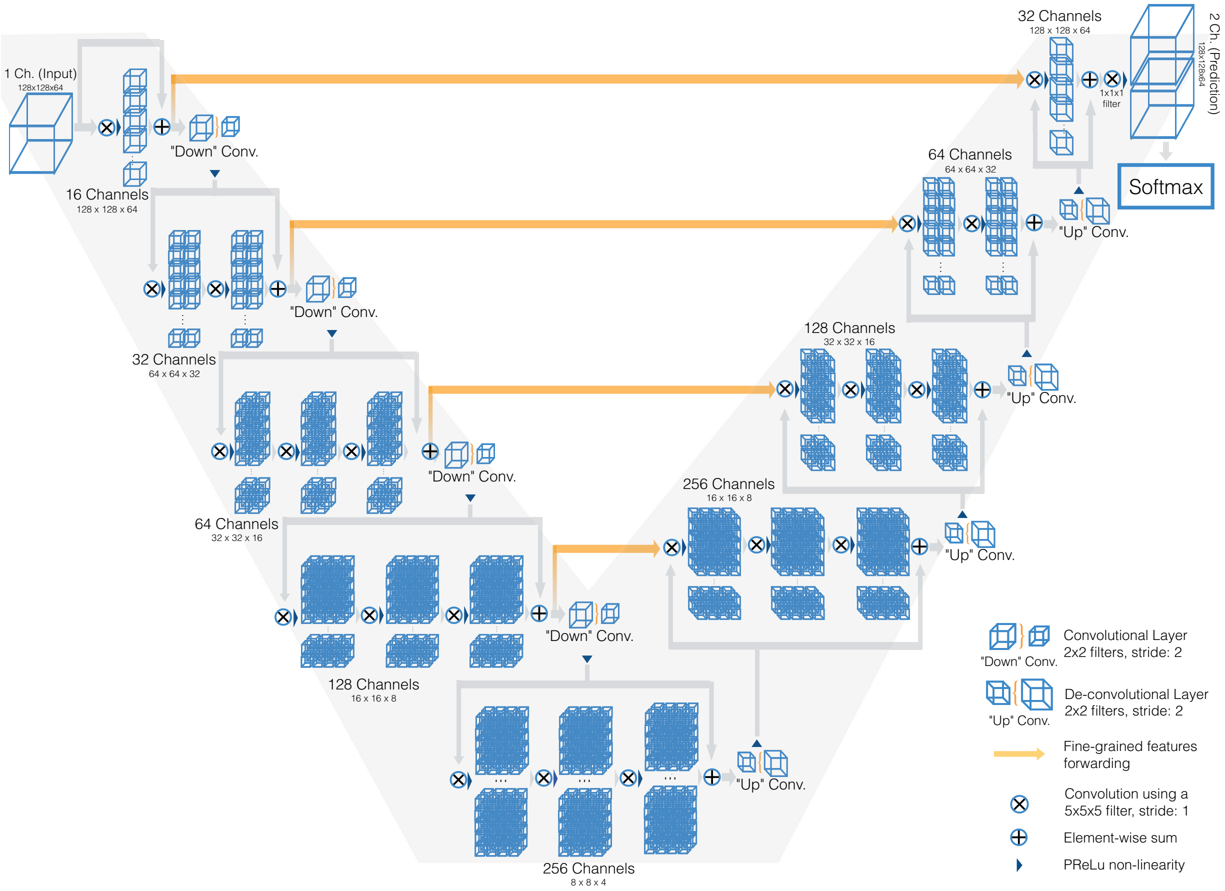

Figure 2. Schematic representation of our network architecture. Our custom implementation of Caffe [8] processes 3D data by performing volumetric convolutions. Best viewed in electronic format.

the time required by a similar network that does not learn residual functions.

The convolutions performed in each stage use volumetric kernels having size $5\times5\times5$ voxels. As the data proceeds through different stages along the compression path, its resolution is reduced. This is performed through convolution with $2 \times 2 \times 2$ voxels wide kernels applied with stride 2 (Figure 3). Since the second operation extracts features by considering only non-overlapping $2\times2\times2$ volume patches, the size of the resulting feature maps is halved. This strategy serves a similar purpose as pooling layers which, motivated by [20] and other works discouraging the use of max-pooling operations in CNNs, have been replaced in our approach by convolutional ones. Moreover, since the number of feature channels doubles at each stage of the compression path of the V-Net, and due to the formulation of the model as a residual network, we resort to these convolution operations to double the number of feature maps as we reduce their resolution. PReLu non linearities [6] are applied throughout the network.

Replacing pooling operations with convolutional ones also results in networks that, depending on the specific implementation, can have a smaller memory footprint during training. This is due to the fact that switches, which map the output of pooling layers back to their inputs, do not need to be stored for back-propagation. In particular, this can be analysed and better understood [23] when applying only de-convolutions instead of un-pooling operations.

Downsampling allows us to reduce the size of the signal presented as input and to increase the receptive field of the features being computed in subsequent network layers. Each of the stages of the left part of the network, computes a number of features which is two times higher than the one of the previous layer.

The right portion of the network extracts features and expands the spatial support of the lower resolution feature maps in order to gather and assemble the necessary information to output a two channel volumetric segmentation. The two feature maps computed by the very last convolutional layer, having $1 \times 1 \times 1$ kernel size and producing outputs of the same size as the input volume, are converted to probabilistic segmentations of the foreground and background regions by applying soft-max voxelwise. After each stage of the right portion of the CNN, a de-convolution operation is employed in order increase the size of the inputs (Figure 3) followed by one to three convolutional layers involving half the number of $5 \times 5 \times 5$ kernels employed in the previous layer. Similar to the left part of the network, we resort to learn residual functions in the convolutional stages of the right part as well.

Similarly to [17], we forward the features extracted from early stages of the left part of the CNN to the right part. This is schematically represented in Figure 2 by horizontal connections. In this way we gather fine grained detail that would be otherwise lost in the compression path and we improve the quality of the final contour prediction. We also observed that these connections improve the convergence time of the model.

We report in Table 4 the receptive fields of each network layer, showing the fact that the innermost portion of our CNN already captures the content of the whole input volume. We believe that this characteristic is important during segmentation of poorly visible anatomy: the features computed in the deepest layer perceive the whole anatomy of interest at once, since they are computed from data having a spatial support much larger than the typical size of the anatomy we seek to delineate, and therefore impose global constraints.

4. Dice loss layer

The network predictions, which consist of two volumes having the same resolution as the original input data, are processed through a soft-max layer which outputs the probability of each voxel to belong to foreground and to background. In medical volumes such as the ones we are processing in this work, it is not uncommon that the anatomy of interest occupies only a very small region of the scan. This often causes the learning process to get trapped in local minima of the loss function yielding a network whose predictions are strongly biased towards background. As a result the foreground region is often missing or only partially detected. Several previous approaches resorted to loss functions based on sample re-weighting where foreground regions are given more importance than background ones during learning [17]. In this work, we propose a novel objective function based on Dice coefficient, a quantity ranging between 0 and 1, which we aim to maximise. The Dice coefficient D between two binary volumes can be written as

$$D = \frac{2 \sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2}$$

where the sums run over the N voxels, of the predicted binary segmentation volume $p_i \in P$ and the ground truth binary volume $g_i \in G$. This formulation of Dice can be differentiated with respect to the j-th voxel of the prediction, yielding the gradient:

$$\frac{\partial D}{\partial p_j} = 2 \left[\frac{g_j \left(\sum_i^N p_i^2 + \sum_i^N g_i^2 \right) - 2 p_j \left(\sum_i^N p_i g_i \right)}{\left(\sum_i^N p_i^2 + \sum_i^N g_i^2 \right)^2} \right]$$

Using this formulation, we do not need to establish the right balance between foreground and background voxels, e.g. by assigning loss weights to samples of different classes such as in [17]. In fact, we obtain results that we experimentally observed are much better than the ones computed through the same network trained optimising a multinomial logistic loss with sample re-weighting (Fig. 6).

Layer	Input Size	Receptive Field
L-Stage 1	128	$5 \times 5 \times 5$
L-Stage 2	64	$22 \times 22 \times 22$
L-Stage 3	32	$72 \times 72 \times 72$
L-Stage 4	16	$172 \times 172 \times 172$
L-Stage 5	8	$372 \times 372 \times 372$
R-Stage 4	16	$476 \times 476 \times 476$
R-Stage 3	32	$528 \times 528 \times 528$
R-Stage 2	64	$546 \times 546 \times 546$
R-Stage 1	128	$551 \times 551 \times 551$
Output	128	$551 \times 551 \times 551$

Table 1. Theoretical receptive field of the $3 \times 3 \times 3$ convolutional layers of the network.

4.1. Training

Our CNN is trained end-to-end on a dataset of prostate scans in MRI. An example of the typical content of such volumes is shown in Figure 1. All the volumes processed by the network have fixed size of $128 \times 128 \times 64$ voxels and a spatial resolution of $1 \times 1 \times 1.5$ millimeters.

Annotated medical volumes are not easily obtainable due to the high cost associated with one or more experts manually tracing a reliable ground truth annotation. In this work we found necessary to augment the original training dataset in order to obtain robustness and increased precision on the test dataset.

During every training iteration, we fed as input to the network randomly deformed versions of the training images by using a dense deformation field obtained through a $2 \times 2 \times 2$ grid of control-points and B-spline interpolation. These augmentations were performed ”on-the-fly”, prior to each optimisation iteration, in order to alleviate the otherwise excessive storage requirements. Additionally, we vary the intensities of the data during training to simulate the variety of data appearance from the scanner. To this end, we use histogram matching to adapt the intensity distributions of the training volumes used in each iteration to the ones of other randomly chosen scans belonging to the dataset.

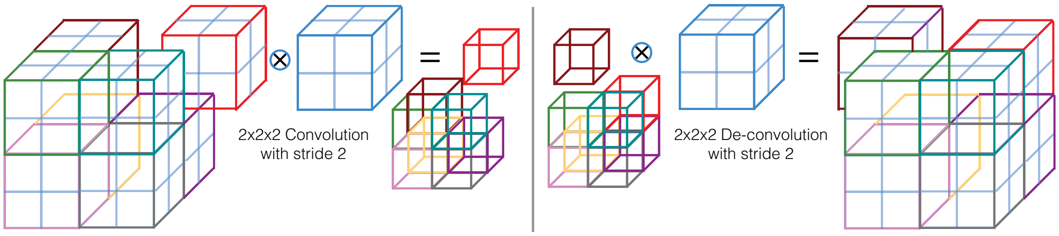

Figure 3. Convolutions with appropriate stride can be used to reduce the size of the data. Conversely, de-convolutions increase the data size by projecting each input voxel to a bigger region through the kernel.

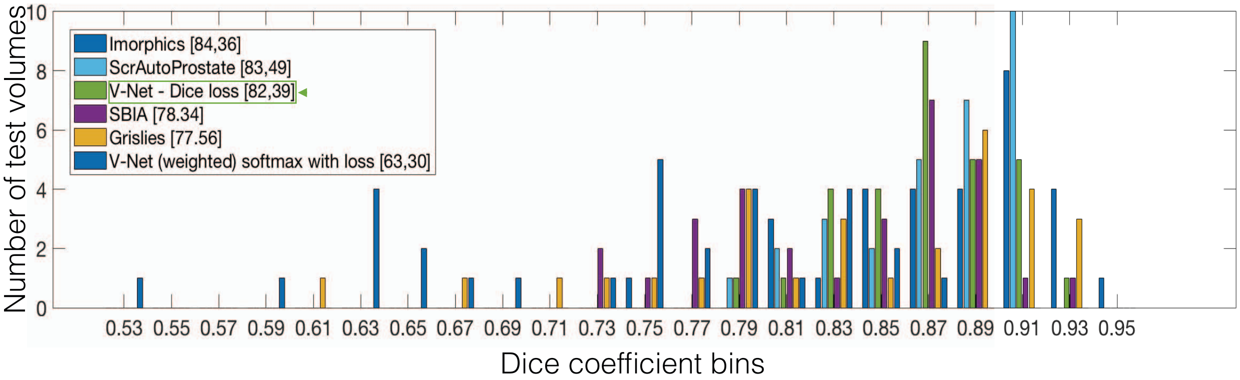

Figure 4. Distribution of volumes with respect to the Dice coefficient achieved during segmentation.

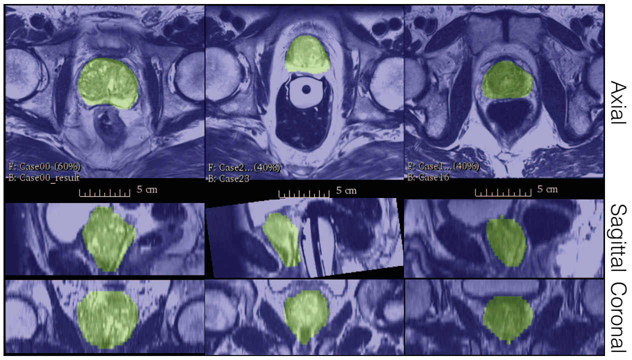

Figure 5. Qualitative results on the PROMISE 2012 dataset [10].

4.2. Testing

A previously unseen MRI volume can be segmented by processing it in a feed-forward manner through the network. The output of the last convolutional layer, after softmax, consists of a probability map for background and foreground. The voxels having higher probability (> 0.5) to belong to the foreground than to the background are considered part of the anatomy.

5. Results

We trained our method on 50 MRI volumes, and the relative manual ground truth annotation, obtained from the ”PROMISE2012” challenge dataset [10]. This dataset contains medical data acquired in different hospitals, using different equipment and different acquisition protocols. The data in this dataset is representative of the clinical variability and challenges encountered in clinical settings. As previously stated we massively augmented this dataset through random transformation performed in each training iteration, for each mini-batch fed to the network. The mini-batches used in our implementation contained two volumes each, mainly due to the high memory requirement of the model during training. We used a momentum of 0.99 and an initial learning rate of 0.0001 which decreases by one order of magnitude every 25K iterations.

We tested V-Net on 30 MRI volumes depicting prostate whose ground truth annotation was secret. All the results reported in this section of the paper were obtained directly from the organisers of the challenge after submitting the segmentation obtained through our approach. The test set was representative of the clinical variability encountered in prostate scans in real clinical settings [10].

We evaluated the approach performance in terms of Dice overlap and Hausdorff distance between the predicted delineation and the ground truth annotation as well as the obtained challenge score, as computed by the organisers of ”PROMISE 2012” [10] (cf. Table 5, Fig. 4).

Our implementation[2] was realised in python, using a custom version of the Caffe[3] [8] framework which was enabled

[2] Implementation available at https://github.com/faustomilletari/VNet

[3] Implementation available at https://github.com/faustomilletari/3D-Caffe

Algorithm	Avg. Dice	Avg. Hausdorff distance	Score on challenge task	Speed
V-Net + Dice-based loss	0.869 ± 0.033	5.71 ± 1.20 mm	82.39	1 sec.
V-Net + mult. logistic loss	0.739 ± 0.088	10.55 ± 5.38 mm	63.30	1 sec.
Imorphics [22]	0.879 ± 0.044	5.935 ± 2.14 mm	84.36	8 min.
ScrAutoProstate	0.874 ± 0.036	5.58 ± 1.49 mm	83.49	1 sec.
SBIA	0.835 ± 0.055	7.73 ± 2.68 mm	78.33	–
Grislies	0.834 ± 0.082	7.90 ± 3.82 mm	77.55	7 min.

Table 2. Quantitative comparison between the proposed approach and the current best results on the PROMISE 2012 challenge dataset.

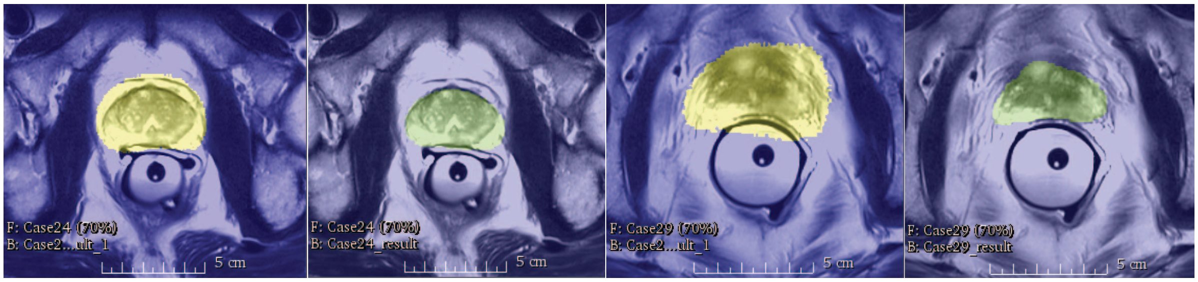

Figure 6. Qualitative comparison between the results obtained using the Dice coefficient based loss (green) and re-weighted soft-max with loss (yellow).

to perform volumetric convolutions via CuDNN v3. All trainings and experiments ran on a standard workstation (64 GB RAM, 3.30GHz Intel® Core™ i7-5820K CPU, NVidia GTX 1080 with 8 GB VRAM). Model training ran for 48 hours, or 30K iterations circa, while segmentation of a previously unseen volume took circa 1 second. Datasets were first normalised using the N4 bias field correction function [21] and then resampled to a common resolution of $1 \times 1 \times 1.5$ mm. We applied random deformations to the scans used for training by varying the position of the control points with random quantities obtained from gaussian distribution with zero mean and 15 voxels standard deviation. Qualitative results can be seen in Fig. 5.

6. Conclusion

We presented and approach based on a volumetric convolutional neural network that performs segmentation of MRI prostate volumes in a fast and accurate manner. We introduced a novel objective function that we optimise during training based on the Dice overlap coefficient between the predicted segmentation and the ground truth annotation. Our Dice loss layer does not need sample re-weighting when the amount of background and foreground pixels is strongly unbalanced and is indicated for binary segmentation tasks. Although we inspired our architecture to the one proposed in [17], we divided it into stages that learn residuals and, as empirically observed, improve both results and convergence time. Future works will aim at segmenting volumes containing multiple regions in other modalities such as ultrasound and at higher resolutions by splitting the network over multiple GPUs.

7. Acknowledgement

We would like to acknowledge NVidia corporation, that donated a Tesla K40 GPU to our group enabling this research, Dr. Geert Litjens who dedicated some of his time to evaluate our results against the ground truth of the PROMISE 2012 dataset and Ms. Iro Laina for her support to this project.

References

[1] Bernard, O., Bosch, J., Heyde, B., Alessandrini, M., Barbosa, D., Camarasu-Pop, S., Cervenansky, F., Valette, S., Mirea, O., Bernier, M., et al.: Standardized evaluation system for left ventricular segmentation algorithms in 3D echocardiography. Medical Imaging, IEEE Transactions on (2015) 1

[2] Cha, K.H., Hadjiiski, L., Samala, R.K., Chan, H.P., Caoili, E.M., Cohan, R.H.: Urinary bladder segmentation in CT urography using deep-learning convolutional neural network and level sets. Medical Physics 43(4), 1882–1896 (2016) 2

[3] Cicek, O., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. arXiv preprint arXiv:1606.06650 (2016) 2

[4] Crum, W.R., Camara, O., Hill, D.L.G.: Generalized overlap measures for evaluation and validation in

medical image analysis. IEEE Transactions on Medical Imaging 25(11), 1451–1461 (2006) 2

[5] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015) 2

[6] He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1026–1034 (2015) 3

[7] Huyskens, D.P., Maingon, P., Vanuytsel, L., Remouchamps, V., Roques, T., Dubray, B., Haas, B., Kunz, P., Coradi, T., Bühlman, R., et al.: A qualitative and a quantitative analysis of an auto-segmentation module for prostate cancer. Radiotherapy and Oncology 90(3), 337–345 (2009) 1

[8] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014) 3, 5

[9] Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation. arXiv preprint arXiv:1603.05959 (2016) 2

[10] Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., et al.: Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Medical image analysis 18(2), 359–373 (2014) 2, 5

[11] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440 (2015) 2

[12] Milletari, F., Ahmadi, S.A., Kroll, C., Plate, A., Rozanski, V., Maiostre, J., Levin, J., Dietrich, O., Ertl-Wagner, B., Bötzel, K., et al.: Hough-CNN: Deep Learning for Segmentation of Deep Brain Regions in MRI and Ultrasound. arXiv preprint arXiv:1601.07014 (2016) 2

[13] Moradi, M., Mousavi, P., Boag, A.H., Sauerbrei, E.E., Siemens, D.R., Abolmaesumi, P.: Augmenting detection of prostate cancer in transrectal ultrasound images using SVM and RF time series. Biomedical Engineering, IEEE Transactions on 56(9), 2214–2224 (2009) 1

[14] Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1520–1528 (2015) 2

[15] Porter, C.R., Crawford, E.D.: Combining artificial neural networks and transrectal ultrasound in the diagnosis of prostate cancer. Oncology (Williston Park, NY) 17(10), 1395–9 (2003) 1

[16] Roehrborn, C.G., Boyle, P., Bergner, D., Gray, T., Gittelman, M., Shown, T., Melman, A., Bracken, R.B., deVere White, R., Taylor, A., et al.: Serum prostate-specific antigen and prostate volume predict long-term changes in symptoms and flow rate: results of a four-year, randomized trial comparing finasteride versus placebo. Urology 54(4), 662–669 (1999) 1

[17] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, pp. 234–241. Springer (2015) 2, 4, 6

[18] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 (2013) 2

[19] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 2

[20] Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806 (2014) 3

[21] Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4ITK: improved N3 bias correction. Medical Imaging, IEEE Transactions on 29(6), 1310–1320 (2010) 6

[22] Vincent, G., Guillard, G., Bowes, M.: Fully automatic segmentation of the prostate using active appearance models. MICCAI Grand Challenge PROMISE 2012 (2012) 6

[23] Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Computer vision–ECCV 2014, pp. 818–833. Springer (2014) 1, 3

[24] Zettinig, O., Shah, A., Hennersperger, C., Eiber, M., Kroll, C., Kübler, H., Maurer, T., Milletari, F., Rackerseder, J., zu Berge, C.S., et al.: Multimodal image-guided prostate fusion biopsy based on automatic deformable registration. Int. J. of Comp. Ass. Rad. Surg. 10(12), 1997–2007 (2015) 1