LONG-RANGE 3D SELF-ATTENTION FOR MRI PROSTATE SEGMENTATION

Federico Pollastri, Marco Cipriano, Federico Bolelli, Costantino Grana

Department of Engineering "Enzo Ferrari," University of Modena and Reggio Emilia, Modena, Italy

ABSTRACT

The problem of prostate segmentation from Magnetic Resonance Imaging (MRI) is an intense research area, due to the increased use of MRI in the diagnosis and treatment planning of prostate cancer. The lack of clear boundaries and huge variation of texture and shapes between patients makes the task very challenging, and the 3D nature of the data makes 2D segmentation algorithms suboptimal for the task. With this paper, we propose a novel architecture to fill the gap between the most recent advances in 2D computer vision and 3D semantic segmentation. In particular, the designed model retrieves multi-scale 3D features with dilated convolutions and makes use of a self-attention transformer to gain a global field of view. The proposed Long-Range 3D Self-Attention block allows the convolutional neural network to build significant features by merging together contextual information collected at various scales. Experimental results show that the proposed method improves the state-of-the-art segmentation accuracy on MRI prostate segmentation.

***Index Terms*—** Prostate MRI, 3D Segmentation, Self-Attention

1. INTRODUCTION

Prostate cancer is the second most commonly diagnosed cancer and the sixth leading cause of cancer death among men worldwide [1]. It represents the third predicted cause of cancer deaths in EU [2] and the second leading cause of cancer death among men in the United States [3]. Although prostate cancer incidence and mortality rates have been proved either to be on the decline or to have stabilized in many countries in recent years, its burden is expected to increase due to the growth and aging of the population [1]: a fast and accurate diagnosis and treatment planning play a fundamental role. Depending on the risk of recurrence and extension of the disease, prostate cancer treatment may consist of radiotherapy or in the surgical removal of the prostate gland [3]. In both cases, the identification and isolation of the gland volume is a prerequisite of the treatment planning, to localize boundaries for external beam radiation therapy or to initialize multi-modal registration algorithms [4]. This is usually performed through Magnetic Resonance Imaging (MRI), a non invasive modality representing the gold standard for prostate imaging due to its superior soft tissue contrast with respect to Computed Tomography (CT) [5].

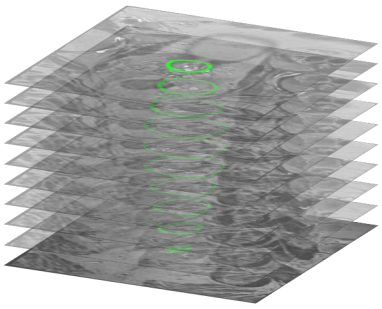

Fig. 1. 2D slices from 3D prostate MRI scan. Annotations are depicted in green. Best viewed in color.

However, performing a manual segmentation of 3D MRI is extremely costly, time consuming, and subject to inter and intra-observer variations. Reliable automated segmentation algorithms could mitigate the aforementioned limitations (Fig. 1). In order to encourage research advances in automatic prostate segmentation, the Prostate MR Image Segmentation PROMISE12 challenge was raised in 2012 [4]. Contextually with the opening of the challenge, a set of 50 segmented MRI from 4 different centers were publicly released and, due to the critical relevance of the task, those data have found applicability even in recent works [6]. Nevertheless, modern deep learning models require significant amounts of data [7, 8, 9, 10] which is not provided in this 2012 public dataset. This paper is build on a a new dataset, released in 2020, which contains a huge amount of annotated prostate biopsies and constitutes a potential milestone for deep learning applied to MRI prostate segmentation.

The main contributions of this work can be summed up as follows:

- A novel 3D segmentation Convolutional Neural Network (CNN) is designed, the architecture of which is motivated in Section 2 and described in detail in Section 3.
- A data pre-processing algorithm is presented in Section 4, to extrapolate a *deep learning-functional* dataset from a publicly available collection of MRI scans.
- In order to grant reproducibility and encourage further advance on this subject, the code for both the data refinement

process and the proposed model is publicly available. Experimental results are presented in Section 5, and in Section 6 conclusions are drawn.

2. RELATED WORKS

U-Net [11] and U-Net3D [12] are two of the most employed architectures in medical imaging. As a matter of fact, Zhu et al. recently approached the semantic segmentation of prostate MRIs using U-Net as a baseline for further improvements [13]. Nowadays, V-Net [14] architecture, almost identical to that of U-Net3D, is considered a state-of-the-art CNN in 3D prostate segmentation [15].
One of the major features introduced by U-Net are the long term connections which became a cornerstone for any semantic segmentation model. Yu et al. aimed to segment the prostate with an higher accuracy by exploiting both long and short term residual connections [6], underlining the relevance of a proper use of low and high level features when dealing with segmentation of medical data. Two years later, in 2019, residual blocks were adopted by Chen et al. to build the backbone of Med3D [15], a segmentation model designed to tackle several distinct medical imaging tasks.

More recently, 2D segmentation models have been improved thanks to multi-scale feature extractors [16, 17]. As an example, the DeepLab architecture [18] introduced a multi-scale architecture to exploit dilated convolutions and gain a wider view of the feature maps, which are merged with low level features using one single long term skip connection. Moreover, several recent proposals addressed transformers as a novel framework which inherits non-local operations [19], the outcome of which, at a given input position, results in a weighted sum of the features at all positions. Transformers were introduced for Natural Language Processing [20] but found numerous applications in a huge range of computer vision tasks [21, 22].

Despite the most modern improvements, 2D convolutions do not take into account the entire spatial information of 3D data. We therefore build a 3D multi-scale neural network suitable for three-dimensional environments, employing ResNet-3D [23] as a backbone to benefit from the effects of short-term residual connections while extracting features. Subsequently, series of multi-spacial features at different scales are computed by means of dilated convolutional filters and additionally enhanced by means of Visual Transformers (ViTs). Finally, feature maps are fed to a decoder, thus reestablishing the input resolution. Instead of feeding a ViT with raw images by splitting them into patches, we rather apply self-attention to our sets of feature maps, and therefore provide the network with a greater non-local awareness of all the extracted spatial information before the final layers. Our novel architecture makes no use of long-term skip connections, which are very expensive in terms of both training time and memory footprint, and would make the proposed method not feasible for training on the most recent hardware.

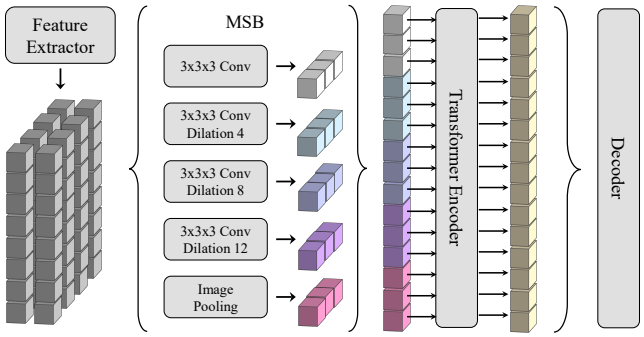

Fig. 2. Schema of the proposed model. The feature extractor generates 512 maps of dimension $14 \times 18 \times 18$, which are fed to the five layers in the Multi-Scale Block (MSB). The first 4 layers capitalize on dilated convolutions to compute contextual information at different scales, whereas the image pooling layer computes image-level features. Layers inside the MSB generate groups of 64 feature maps, each depicted with a different color within the Figure. The 320 features are then enhanced by the transformer encoder (detailed in Fig. 3), before being fed into the decoder. Best viewed in color.

3. PROPOSED METHOD

The proposed 3D architecture can be divided in three main components: the feature extraction CNN (Table 1), the Long-Range 3D Self-Attention Block (Fig. 2), and the decoder (Table 2). Every convolutional layer in the model –but the last one of the decoder, named *LastC*– is followed by 3D batch normalization [24] and a ReLu activation function.

3.1. Feature Extractor.

Rich semantic features represent a key element for every computer vision task, and the ability of Convolutional Neural Networks to autonomously learn how to extract them is the main reason behind their groundbreaking rise. We thus exploit a pre-trained 3D Resnet-18 [23] to compute meaningful features across all of the three dimensions that characterize our data, and slightly modify its architecture to increase the spatial resolution of the output. Following the guidelines defined in the paper, the stride of the first convolution of the first block of *Layer2* and *Layer3* is set to $2 \times 2 \times 2$. Contrarily to the original architecture, we apply every convolution in *Layer4* with no stride, and set the dilation of the second block convolutions to $2 \times 2 \times 2$, thus obtaining 512 feature maps as the output of our feature extractor, each of size $14 \times 18 \times 18$. This change is motivated by the fact that the original CNN was designed for a classification task, whereas semantic segmentation neural networks benefit from larger feature maps resolution [25].

Table 1. Detailed description of the feature extractor.

Layer	Output	Block
Center Crop	$56 \times 144 \times 144$	
Stem	$56 \times 72 \times 72$	$3 \times 7 \times 7, 64$
Layer1	$56 \times 72 \times 72$	$\begin{bmatrix} 3 \times 3 \times 3, & 64 \\ 3 \times 3 \times 3, & 64 \end{bmatrix} \times 2$
Layer2	$28 \times 36 \times 36$	$\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix} \times 2$
Layer3	$14 \times 18 \times 18$	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times 2$
Layer4	$14 \times 18 \times 18$	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times 2$
LR3DSA	$14 \times 18 \times 18$	Fig. 2

Table 2. Detailed description of the decoder.

Layer	Output	Block
LR3DSA	$14 \times 18 \times 18$	Fig. 2
Interpolate1	$56 \times 72 \times 72$	$4 \times 4 \times 4$
Conv	$56 \times 72 \times 72$	$\begin{bmatrix} 3 \times 3 \times 3, 320 \\ 3 \times 3 \times 3, 320 \end{bmatrix} \times 1$
LastC	$56 \times 72 \times 72$	$1 \times 1 \times 1, 1$
Interpolate2	$56 \times 144 \times 144$	$1 \times 2 \times 2$
Activation	$56 \times 144 \times 144$	sigmoid

3.2. Multi-Scale Self-Attention Block.

With the purpose of obtaining multi-scale contextual information, we design a Multi-Scale Block (MSB) as the 3D extension of the 2D Atrous Spatial Pyramid Pooling [18]. Our MSB is composed of four different dilated convolutions and one average pooling layer, which are all fed with the same extracted feature maps. Given the dimensions of the maps computed by the feature extractor, dilation hyperparameters are set to 1, 4, 8, and 12, in order to take into account the relationship between voxels at increasing distance when generating the multi-scale features. The average pooling layer serves to yield image-level features, and is followed by a $1 \times 1 \times 1$ convolution to reduce the number of computed features; each one of the five layers in the MSB outputs 64 feature maps of dimension $14 \times 18 \times 18$.

The 320 feature maps are then flattened and fed to a transformer encoder described in Fig. 3. By collectively evaluating feature maps computed at different scales, the transformer is able to integrate both short-range and long-range features, gaining a global view of the image which is very beneficial for 3D prostate MRI scans and comes at a limited computational cost. The proposed self-attention module does not need positional encoding, since it is fed unordered features.

3.3. Decoder.

Finally, the original spatial dimensions are reestablished through the decoder. A first interpolation increases the size of the feature maps to $56 \times 72 \times 72$, fully restoring the resolution across the z axis, while leaving the dimensions halved across the other two. The larger feature maps are elaborated by two cascaded $3 \times 3 \times 3$ convolutions, before a single $1 \times 1 \times 1$ filters obtain the single channel prediction from the 320 input features. The last interpolation enlarges the dimensions of the image along x and y, yielding a prediction with the same resolution of the input volume.

4. DATASET

The proposed architecture is trained and tested on data gathered from the Prostate-MRI-US-Biopsy dataset [26], which was collected by the National Cancer Institute in Maryland (USA) and contains of $2\,799$ studies over $1\,151$ patients. Out of the $2\,799$ studies, only $1\,017$ provide 3D volumes stored as DICOM files, coupled with the corresponding manual prostate segmentation as Standard Triangulation Language (STL) files. We discard 103 studies with z resolution different from 60, thus obtaining a dateset of 3D uniquely axial scans with a fixed number of slices. Our final dataset is composed of 911 volumes with shapes $60 \times 256 \times 256$ or $60 \times 512 \times 512$, which are split into a training, validation and test sets with respectively 711, 40 and 160 volumes.

STL annotation files describe each prostate as a series of mesh, namely sets of triangles mapping its surface. Vertex values from each triangle are firstly rotated and scaled with respect to the patient acquisition setting, volume voxels which intersect the triangles are then detected through a voxelization algorithm. Binary segmentation masks are finally stored as `.npy` (numpy) files.

Our refined version of the Prostate-MRI-US-Biopsy

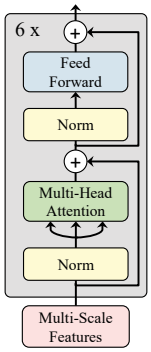

Fig. 3. Transformer encoder inspired by [20].

Table 3. Performance of the proposed method, compared against both 3D and 2D competitors in three different dataset set-ups.

Method	Prostate-MRI-US-Biopsy				PROMISE12				Fine-Tuned for PROMISE12			
	Volume IoU	Slice IoU	Volume DICE	Slice DICE	Volume IoU	Slice IoU	Volume DICE	Slice DICE	Volume IoU	Slice IoU	Volume DICE	Slice DICE
Ours	0.846	0.859	0.916	0.895	0.716	0.726	0.834	0.775	0.785	0.807	0.880	0.847
V-Net	0.822	0.840	0.901	0.880	0.390	0.463	0.551	0.534	0.692	0.761	0.815	0.811
Med3D	0.822	0.840	0.901	0.880	0.653	0.714	0.787	0.762	0.736	0.776	0.847	0.821
U-Net3D	0.822	0.840	0.901	0.880	0.482	0.519	0.635	0.584	0.704	0.740	0.824	0.790
DeepLabv3+	0.826	0.841	0.904	0.880	0.701	0.735	0.821	0.782	0.759	0.803	0.862	0.848
U-Net	0.776	0.810	0.871	0.855	0.699	0.779	0.820	0.822	0.763	0.814	0.865	0.857

dataset is composed of volumes with very homogeneous resolutions, and the proposed architecture is tailored to such sizes. On the other hand, uniform resolution is not something that can be guaranteed when working with prostate MRI scans. Therefore, we explore the effectiveness of the proposed method when applied to data with different resolutions by means of the PROMISE12 dataset [4], which contains 50 public annotated scans obtained from different domains, and is characterized by extremely variable resolutions that range from $18 \times 256 \times 256$ to $54 \times 512 \times 512$. We split this public dataset into a training set of 40 scans, and a test set of 10 scans.

5. EXPERIMENTAL RESULTS

The performance are evaluated through the two metrics Intersection over Union (IoU) and Dice Coefficient, computed both per-volume and per-slice. Intuitively, when evaluating a volume metric every voxel within the whole scan has the same importance, whereas slice metrics are the mean of the values computed per-slice. Slice metrics are more punishing towards a wrong prediction on an empty slice, since the score is 1 for correct predictions, and 0 when even only 1 voxel is labeled as foreground. The results are compared against several segmentation CNNs, divided into 3D architectures [12, 14, 15] and 2D architectures [11, 25].

Table 3 shows that, on the refined Prostate-MRI-US-Biopsy dataset, the proposed method outperforms every competitor for each one of the computed metrics. The 4 columns under the PROMISE12 section display the results obtained by the networks when trained and tested using only the PROMISE12 dataset and, finally, the last 4 columns evaluate the models when pre-trained with the Prostate-MRI-US-Biopsy and fine-tuned for the PROMISE12 dataset. Before being fed to 3D neural networks, every scan is resized to $60 \times 256 \times 256$, and sub-volumes of dimensions $56 \times 144 \times 144$ are obtained by center-cropping the entire volumes. However, no rescaling along the z axis is required for 2D neural networks. Thus, when processing volumes with variable resolutions, slice metrics favor 2D architectures over 3D ones, since the resampling operations along the z axis can generate minor errors in slices with no foreground. The volume metrics, however, demonstrate that the proposed method always performs prostate segmentation with better accuracy overall. Moreover, pre-training a model with the Prostate-MRI-US-Biopsy dataset always improve every segmentation metric for each one of the tested architectures, at least in the analyzed domain (refer to the last four columns of Table 3).

Data augmentation is performed by means of horizontal flipping and rotations of random angles that range from -8 to 8 degrees. CNNs are trained to optimize a joined loss obtained by summing a weighted cross-entropy loss and a Jaccard loss [27], the Stochastic Gradient Descent optimizer is employed, and the initial learning rate of 0.1 is gradually decreased by means of a plateau scheduler.

6. CONCLUSIONS

This paper presents a novel 3D CNN to address prostate segmentation in MRI scans. The designed Long-Range 3D Self-Attention block is able to elaborate global features combining information collected at various scales, by merging the properties of dilated convolutions and self-attention. The proposed architecture can be easily trained in an end-to-end fashion thanks to its simple yet effective nature and, given the reduced number of convolutional filters and the complete absence of long-term skip connections, a limited amount of resources is needed to complete the training process (1 GPU for 25 hours). Experimental results showcase that the proposed method outperforms its competitors in MRI prostate segmentation.

Furthermore, in this work we designed a series of pre-processing steps to refine a collection of publicly available prostate MRI scans and build a *deep learning-functional* dataset. Experimental results demonstrate that the proposed refined dataset is beneficial for the prostate segmentation research field, when employed for CNN pre-training. The publicly available code[1] can be used to reassemble the employed dataset, split the available data in the same partitions used in the experiments (test, validation, and training sets), and compare the proposed method against popular state-of-the-art models and newly developed architectures.

[1] `github.com/PollastriFederico/3D-self-attention`

7. REFERENCES

[1] M. B. Culp, I. Soerjomataram, J. A. Efstathiou, et al., "Recent Global Patterns in Prostate Cancer Incidence and Mortality Rates," *European Urology*, vol. 77, no. 1, pp. 38–52, 2020.

[2] G. Carioli, P. Bertuccio, P. Boffetta, et al., "European cancer mortality predictions for the year 2020 with a focus on prostate cancer," *Annals of Oncology*, vol. 31, no. 5, pp. 650–658, 2020.

[3] Y. Lei, X. Dong, Z. Tian, et al., "CT prostate segmentation based on synthetic MRI-aided deep attention fully convolution network," *Medical Physics*, vol. 47, no. 2, pp. 530–540, 2020.

[4] G. Litjens, R. Toth, W. van de Ven, et al., "Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge," *Medical Image Analysis*, vol. 18, no. 2, pp. 359–373, 2014.

[5] B. Wang, Y. Lei, S. Tian, et al., "Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic MRI prostate segmentation," *Medical Physics*, vol. 46, no. 4, pp. 1707–1718, 2019.

[6] L. Yu, X. Yang, H. Chen, et al., "Volumetric ConvNets with Mixed Residual Connections for Automated Prostate Segmentation from 3D MR Images," in *AAAI Conference on Artificial Intelligence*, 2017, vol. 31.

[7] C. Mercadante, M. Cipriano, F. Bolelli, et al., "A Cone Beam Computed Tomography Annotation Tool for Automatic Detection of the Inferior Alveolar Nerve Canal," in *VISAPP*. 2021, vol. 4, pp. 724–731, SciTePress.

[8] F. Pollastri, F. Bolelli, R. Paredes, and C. Grana, "Improving Skin Lesion Segmentation with Generative Adversarial Networks," in *CBMS*, 2018, pp. 442–443.

[9] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "SAM: Pushing the Limits of Saliency Prediction Models," in *CVPR Workshops*, 2018.

[10] R. Bigazzi, F. Landi, S. Cascianelli, L. Baraldi, M. Cornia, and R. Cucchiara, "Focus on Impact: Indoor Exploration with Intrinsic Motivation," *IEEE Robotics and Automation Letters*, 2022.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional Networks for Biomedical Image Segmentation," in *MICCAI*, 2015, pp. 234–241.

[12] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, et al., "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," in *MICCAI*, 2016, pp. 424–432.

[13] Q. Zhu, B. Du, B. Turkbey, P. L. Choyke, and P. Yan, "Deeply-Supervised CNN for Prostate Segmentation," in *IJCNN*, 2017, pp. 178–184.

[14] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *3DV*, 2016, pp. 565–571.

[15] S. Chen, K. Ma, and Y. Zheng, "Med3D: Transfer learning for 3D Medical Image Analysis," *arXiv*, 2019.

[16] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic Feature Pyramid Networks," in *CVPR*, 2019, pp. 6399–6408.

[17] H. Zhao, J. Shi, X. Qi, et al., "Pyramid Scene Parsing Network," in *CVPR*, 2017, pp. 2881–2890.

[18] L.-C. Chen, G. Papandreou, I. Kokkinos, et al., "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *TPAMI*, vol. 40, pp. 834–848, 2017.

[19] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local Neural Networks," in *CVPR*, 2018, pp. 7794–7803.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All You Need," *arXiv*, 2017.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv*, 2020.

[22] M. Cornia, L. Baraldi, and R. Cucchiara, "SMArT: Training Shallow Memory-aware Transformers for Robotic Explainability," in *ICRA*, 2020.

[23] D. Tran, H. Wang, L. Torresani, et al., "A Closer Look at Spatiotemporal Convolutions for Action Recognition," in *CVPR*, 2018, pp. 6450–6459.

[24] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *PMLR*, 2015, pp. 448–456.

[25] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *ECCV*, 2018.

[26] S. Natarajan, A. Priester, D. Margolis, J. Huang, and L. Marks, " Prostate MRI and Ultrasound With Pathology and Coordinates of Tracked Biopsy (Prostate-MRI-US-Biopsy) [Dataset] ," 2020.

[27] F. Pollastri, F. Bolelli, R. Paredes, and C. Grana, "Augmenting Data with GANs to Segment Melanoma Skin Lesions," *Multimedia Tools and Applications*, vol. 79, no. 21-22, pp. 15575–15592, 2019.