A Lumen Segmentation Method in Ureteroscopy Images based on a Deep Residual U-Net architecture

Jorge F Lazo, Aldo Marzullo, Sara Moccia, Michele Catellani, Benoît Rosa, Francesco Calimeri, Michel de Mathelin, Elena de Momi

▶ **To cite this version:**

Jorge F Lazo, Aldo Marzullo, Sara Moccia, Michele Catellani, Benoît Rosa, et al.. A Lumen Segmentation Method in Ureteroscopy Images based on a Deep Residual U-Net architecture. International Conference on Pattern Recognition (ICPR 2020), Jan 2021, Milan (virtuel), Italy. ⟨hal-03051989⟩

HAL Id: hal-03051989

https://hal.science/hal-03051989v1

Submitted on 10 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

A Lumen Segmentation Method in Ureteroscopy Images based on a Deep Residual U-Net architecture

1st Jorge F. Lazo
DEIB, Politecnico di Milano,
ICube, Université de Strasbourg
Milan, Italy
jorgefrancisco.lazo@polimi.it

2rd Aldo Marzullo
Department of Mathematics and Computer Science
University of Calabria
Rende, Italy
marzullo@mat.unical.it

3nd Sara Moccia
Dipartimento di Ingegneria dell'Informazione,
Università Politecnica delle Marche
Milan, Italy
s.moccia@staff.univpm.it

4th Michele Catellani
Department of Urologic Surgery
Istituto Europeo di Oncologia
Milan, Italy
michele.catellani@ieo.it

5th Benoit Rosa
ICube, UMR 7357
CNRS-Université de Strasbourg
Strasbourg, France
b.rosa@unistra.fr

6th Francesco Calimeri
Department of Mathematics and Computer Science
University of Calabria
Rende, Italy
calimeri@mat.unical.it

8th Michel de Mathelin
ICube - UMR CNRS 7357
Université de Strasbourg
Strasbourg, France
demathelin@unistra.fr

7th Elena De Momi
DEIB
Politecnico di Milano
Milan, Italy
elena.demomi@polimi.it

***Abstract*—Ureteroscopy is becoming the first surgical treatment option for the majority of urinary affections. This procedure is performed using an endoscope which provides the surgeon with the visual information necessary to navigate inside the urinary tract. Having in mind the development of surgical assistance systems, that could enhance the performance of surgeon, the task of lumen segmentation is a fundamental part since this is the visual reference which marks the path that the endoscope should follow. This is something that has not been analyzed in ureteroscopy data before. However, this task presents several challenges given the image quality and the conditions itself of ureteroscopy procedures. In this paper, we study the implementation of a Deep Neural Network which exploits the advantage of residual units in an architecture based on U-Net. For the training of these networks, we analyze the use of two different color spaces: gray-scale and RGB data images. We found that training on gray-scale images gives the best results obtaining mean values of Dice Score, Precision, and Recall of 0.73, 0.58, and 0.92 respectively. The results obtained shows that the use of residual U-Net could be a suitable model for further development for a computer-aided system for navigation and guidance through the urinary system.**

This work was supported by the ATLAS project. ATLAS has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 813782.

***Index Terms*—deep learning, ureteroscopy, convolutional neural networks, lumen segmentation,**

I. Introduction

Ureteroscopy is a procedure dedicated to the exploration of the upper urinary tract to perform diagnosis and treatment of different conditions, such as kidney stones or carcinoma of the upper urinary tract. The procedure involves the passage of a flexible or semi-rigid ureteroscope through the urethra and bladder, and up to the ureter to the point where the stones or tumor is located [1]. Urinary calculi may be located anywhere in the urinary tract, but they are more common to appear in the kidneys. In the case of cancer, tumors can be found at any point in the renal pelvis, renal calyces, and ureters, but the most common location is the renal pelvis and renal calyces, being present in about 4-15% of the cases [2]. The total traveling distance from the ureter orifice to the kidneys ranges from 28cm to 34cm and the diameter of the ureter is 5mm in the pelvis [3].

A review of the intraoperative complications in ureteroscopy shows that avulsion, major and minor perforation, mucosal abrasion and stricture are considered

the most common ones [4]. Minor perforation is the one with a higher frequency, appearing with an average of 1.99% of the analyzed procedures, followed by stricture (0.58%), and major perforation is the one with lower probability (0.06% of the cases). Navigation and diagnosis inside the urinary tract are highly dependent upon the operators expertise, and image-related conditions such as the presence of image artifacts, floating debris, the low quality of the video, occlusions in the video, or image noise [5], could add additional challenges for non-experienced operators. Furthermore, the heterogeneity in the anatomical structures, and the different shapes in which the lumen is deformed at different points in the urinary tract adds additional challenge to the task of segmenting the lumen. Some examples which highlight the variability in the deformation, colors and shape of the lumen are shown in Fig. 1.

The task of lumen segmentation is a fundamental part in the development of computer vision methods for surgical assistance, since this is the reference which marks the path that the endoscope should follow. These assistance systems intend to deal with the mentioned challenges by enhancing relevant information to potentially increase the performance of the surgeon and minimize the probability of complications. Besides, the combination of this visual information with the information from some other sensors such as EM sensors or Fiber Bragg Grating sensors, that can provide positional information, is a primary step in the development of computer vision systems suitable for more advanced robot-assisted ureteroscopy [6].

Recent advancements in the field of computer vision applied to biomedical image analysis and particularly, the use of Deep Learning (DL) methods such as Convolutional Neural Networks (CNNs) have shown remarkable results in biomedical image segmentation challenges such as EM ISIBI 2012 [7], the Longitudinal MS Lesion Segmentation Challenge [8] or the Endoscopic Artefact Detection Challenge EAD2019 [5].

In this paper, we propose the application of a variation of the well-known CNN architecture U-Net by using residual blocks [9]. This have shown to have beneficial effects dealing with the vanishing gradient problem [10]. We compare the results of training the network in the same datasets, but in different image modalities, one composed only of gray-scale images and another of RGB images to determine whether the information contained in a color-space of one single channel is enough to perform the task, or a bigger color-space is needed.

The rest of the document is organized as follows: in Sec. II we provide a survey of lumen segmentation methods used in data from other surgical procedures as well as image segmentation methods used in the analysis of ureteroscopy images, Sec. III presents the details of the implementation used in this study, in Sec. IV the results obtained are given and discussed. Finally Sec. VI concludes the paper.

Fig. 1: Sample of images in the dataset collected showing the variability of the hue as well as the shape and location of the tissue which sometimes merges with the tool and the guide-wire.

II. Related Work

Lumen segmentation has been explored in different surgical procedures such as colonoscopy procedures and also in different image modalities such as Intravascular Ultrasound images. We limit the literature research to colon procedures, which not only has a more similar anatomical structure, but the procedure itself is more similar to ureteroscopy regarding the kind of image data and instruments used during the intervention.

The study in [11] implements a intestinal detection method based on the application of a Haar-like features-based algorithm to recognize the lumen. A combination of the algorithm with AdaBoost to select discriminative features is later incorporated in a cascade classifier. The method is sensitive to light conditions, and only works for lumen detection. The method proposed in [12] makes use of a coarse-to-fine version of the mean-shift algorithm to detect and track the lumen. The algorithm runs for several times with different seed points and the detected extreme points are spatially clustered. In [13], two methods are presented, one based on adaptative thresholding for the detection of lumen when is not

completely contracted and an alternate method for the case in which lumen is contracted. The second method is based on texture detection where, first adaptative thresholding is used to locate the minimum gray pixel and then the strong edges (wrinkles) of the folded intestinal wall are detected. Even if this method is capable of detecting the lumen shape no results in term of pixel accuracy are given and it is not optimal when there are considerably amounts of turbid liquid. The main limitation in the implementation of these methods is the fact that they are parameter-sensitive, which means that they cannot generalize very well to unseen data and in general to the inter-patient anatomy variability and imaging conditions. To deal with this more recent approaches have started to make use of Neural Networks-based methods.

A benchmark for segmentation in colonoscopy images is presented in [14]. The authors use a multi-class image segmentation method based on the implementation of a 8-layer Fully Convolutional Network (FCN) and the classes they segment are: background, lumen, polyp and image specularities. Considering the specific case of ureteroscopy images, no study related with the segmentation of lumen has been found. We relate this to the lack of public available datasets. A possible reason for such fact could be difficulty of obtaining this image data and the lack of an automated or semi automated device to perform this procedure.

The review of image segmentation methods in ureteroscopy images, even if it refers to the segmentation of different structures, is relevant for our work given the specific challenges in the analysis of uretersocopy images. Some of these challenges include a low image quality because of breathing motion [15], presence of floating debris from the dusted tissue and the irrigation fluid used to remove it [1].

A kidney stone detection method using ureteroscopy images is proposed in [16] and it is based on a region growing algorithm, this algorithm requires of a seed, a similarity criterion and a stopping criterion. However, the definition of the criterion is not optimal since the position of the seed must be defined by the user (with the requirement that it should be placed in the center of the calculus), and the stopping criterion, which need to be optimized is empirically determined. A method based on Single Shot MultiBox Detector for Ureteral Orifice detection and segmentation was proposed in [17]. The authors use images from resctoscopy, which have a larger field of view, to train their model and images from ureteroscope to validate it. Recently in [18] an implementation for kidney stones segmentation based on the combination of U-Net and a Deformation Vector Field which makes use of consecutive frames to determine the deformation before being fed the regular U-Net architecture has been implemented. Nevertheless, these approaches have the peculiarity that they focus only in a specific and small part of the urinal system and they do not deal with most of the irregularities that appears in uretersocopy images. In our work, we deal not only with the variability of the image quality dependant on the device used in the procedure, the inter-patient anatomical variability, but also the inner-variability of the shape, texture and deformation of the ureter through the urinary tract. Furthermore, we show that the method we propose can be tested on data with the presence of elements that were not present in the data used for training, such as lasers and tumors, and still detect the lumen to some extent.

To deal with this high-variability images we propose the use of Deep CNNs. To the best of our knowledge, CNNs have not been applied to image segmentation of the lumen in ureteroscopy images. We study the implementation of a model based in deep residual networks and compare it with other 2 models in two datasets coming from the same sources, but in different color-space. Furthermore, we implement different data augmentation techniques to improve the results obtained.

III. Proposed Method

Inspired by the paradigm set by residual networks [19] which have shown to achieve state of the art performances in segmentation tasks such as road segmentation from aerial images [20] or nuclei segmentation [21]. We propose the implementation of a segmentation network based on U-Net architecture which makes use of residual blocks. Among the advantages these kind of networks presents there is the reduction in the training time, as result of the residual blocks the convergence of the learning process is faster. Thanks to the short-skip connections, present in this network, the spatial information propagates without degradation and in general, it has been reported that the segmentation with this architectures results in a better delineation of the borders of the figures [10]. We compare the results obtained with the aforementioned network with two other architectures, in this case, a model similar to the standard version of U-Net, which only differs by the addition of Batch Normalization, and FCN-8 network which has been used previously for lumen segmentation.

A. FCN-8

Fully Convolutional Neural Networks (FCNs) are an extension of CNNs to deal with the task of pixel prediction. These networks make use of long skip connections in order to recover the fine-grained spatial information from the down-sampling path by merging it in the up-sampling layers. In this work we implemented a FCN8 architecture as proposed in [14] for endoluminal scene segmentation in endoscopic images.

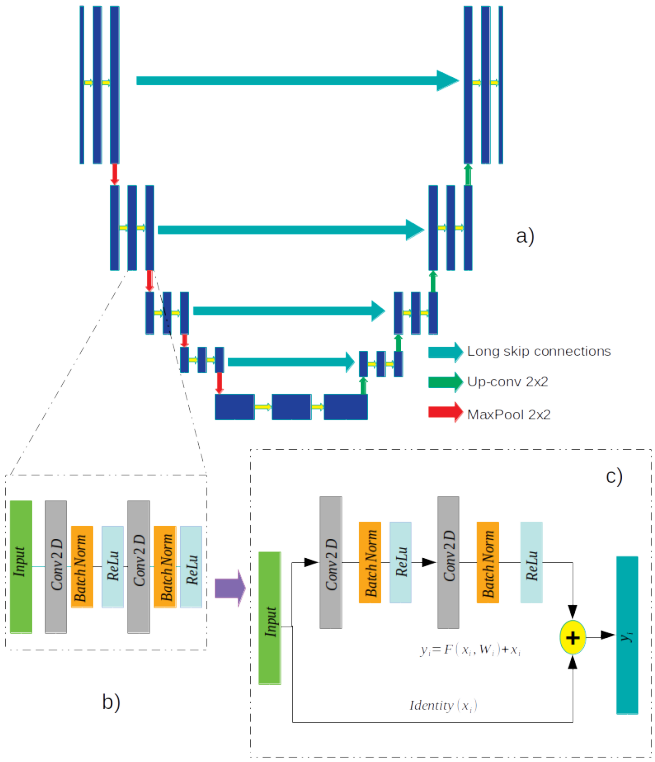

Fig. 2: Comparison of the different convolutional blocks used to build the models. a) General sketch of U-Net architecture. b) Diagram of the convolutional block adding batch normalization c) residual block with short skip-connections.

B. U-Net

U-Net is an architecture which originally was designed for biomedical image segmentation applications [22]. It is composed of a contracting path and an up-sample path which are symmetrical. The convolutional blocks of U-Net uses Convolutional Layers of kernel size 3x3 followed by 2x2 max pooling layers. The number of kernels used in each layer doubles at every block and the output of the contracting stage is connected to the feature maps in the corresponding up-sampling layer. This can ensure that the features learned in the contracting path are used in the up-sampling path for the reconstruction of the segmented images. In the implementation used in this work we added Batch Normalization layers to the output of every convolutional layer. This helps to reduce the amount of covariance shift off the hidden layers and it has the advantage that it has regularization effects, reducing overfitting and allowing the use of higher learning rates [23].

The loss function used in this implementation was based on the Dice similarity coefficient (L_{DSC}) defined as:

$$L_{DSC} = 1 - \frac{2TP}{2TP + FN + FP} \tag{1}$$

where TP is the number of pixels that belong to the lumen, which are correctly segmented, FP is the number of pixels miss-classified as lumen, and FN is the number of pixels which are classified as part of lumen but actually they are not.

C. U-Net based on residual blocks

Residual Blocks were first introduced in the ResNet architecture for image classification task. Their application was later expanded to image segmentation tasks and inspired the construction of networks which take the advantages of this type of convolutional blocks. These kind of networks have proven to achieve performances above the state of the art in both, classification and segmentation tasks. The residual units intend to address the degradation problem by adding a skip connection between the input and the output of each convolutional block, which can be defined as:

$$y_i = h(x_i) + F(x_i, W_i) \tag{2}$$

$$x_{i+1} = f(y_i) \tag{3}$$

where x_i and x_{i+1} are the input and the output respectively of the i^{th} residual block, F is the residual function, $f(y_i)$ is the activation function and $h(x_i)$ is the identity mapping function.

The implementation used in this study consisted of using a basic residual block as defined in [10]. This is composed of a convolution layer followed by batch normalization and a *ReLU* activation function which later was adhered to the output of the conventional convolutional block as depicted in Fig. 2. The loss function used in the network was the same as the used in the previous network and defined by Eq. 1.

D. Performance Metrics

The performance metrics chosen were the DSC, the Precision ($Prec$) and Recall (Rec) which are defined as

$$DSC = 1 - L_{DSC} \tag{4}$$

$$Prec = \frac{TP}{TP + FP} \tag{5}$$

$$Rec = \frac{TP}{TP + FN} \tag{6}$$

TABLE I: Detailed information about the dataset collected. The videos marked with $*$ indicate the videos that were set apart to be used only during testing.

	No. of annotated frames	Image Size (pixels)	Patient No.
Video 1*	7	356x256	1
Video 2*	80	256x266	1
Video 3	462	296x277	2
Video 4	245	256x257	3
Video 5	148	256x257	3
Video 6	168	256x257	3
Video 7	235	256x262	4
Total	1,445	-	-

TABLE II: Resume of the average with its respective standard deviation of the best results obtained with the 5-fold cross validation for each of the models tested in each of the datasets: (RGB dataset white rows, grayscale-dataset gray rows).

Model	DSC	$Prec$	Rec	Acc
U-Net	0.67 ± 0.01	0.57 ± 0.01	0.69 ± 0.03	$\mathbf{0.94 \pm 0.02}$
U-Net	0.55 ± 0.09	0.43 ± 0.11	0.57 ± 0.07	0.87 ± 0.17
Residual U-Net	$\mathbf{0.68 \pm 0.05}$	$\mathbf{0.58 \pm 0.01}$	$\mathbf{0.78 \pm 0.03}$	0.92 ± 0.10
Residual U-Net	0.59 ± 0.15	0.49 ± 0.03	0.63 ± 0.13	0.94 ± 0.01
FCN-8	0.36 ± 0.08	0.25 ± 0.09	0.64 ± 0.01	0.78 ± 0.07
FCN-8	0.23 ± 0.04	0.19 ± 0.03	0.52 ± 0.10	0.64 ± 0.15

E. Dataset

For this study 7 videos were collected. The videos were acquired from the European Institute of Oncology (IEO) at Milan, Italy.All patients gave their informed consent for data collection and use of it for research. The data collection followed the ethical protocol approved by the IEO and in accordance with the Helsinky Declaration. The videos collected are from ureteroscopy procedures targeting upper tract tumor ablation and kidney stone removal. From these videos a different number of frames were extracted as described in Table I and manually annotated. Some samples of the images from the dataset which depict the variability in the images is shown in Fig. 1

A prepossessing stage included cropping of the frames to cut out the black region surrounding the field of view of the endoscopic images, and the conversion to gray-scale for the training and testing in the gray-scale dataset. The former one, was done by applying a canny filter for edge detection and subsequently an elliptical Hough transform algorithm based on [24]. The final shape of the images was set to 256x256 to be consistent with the input layer of the networks used.

F. Training Setting

To handle memory limitations the networks were trained with a mini-batch size strategy. The networks were trained using data augmentation, the operations used for this purpose were rotation in intervals of 90°, 180°, and 270°; horizontal and vertical flips and zooming in and out in a range of $\pm$ 0.02. The learning rate, and mini batch size for each of the models was chosen by trying the different combinations between several possible values of the hyper-parameters and using a 5-fold cross validation strategy with the data from patients 2, 3 and 4. The 5-fold process was repeated independently with both the RGB dataset and the gray-scale dataset. The DSC was set as the evaluation metric to chose the best model with respect to the hyper-parameters. Adam optimization was used during all the trainings. Once the hyper-parameter values were chosen, an additional training process was carried out using these values. This training was done by using all the annotated frames obtained from patients 2, 3 and 4 for training and validation, while the data from patient 1 was used for testing. In total 798 frames were used for training and 417 for validation in this stage. The Kruskal-Wallis test on the DSC was used to determine statistical significance between the different models trained.

The Networks were implemented using *Tensorflow* and *Keras* frameworks in Python trained on a *NVIDIA Quadro M5000* GPU.

IV. Results

The box plots of the DSC, $Prec$ and Rec are depicted in Fig. 4 for each of the networks tested. From these results is possible to see that Residual U-Net has the best performance overall in DSC, $Prec$ and Rec with average values of 0.73, 0.58, and 0.92 respectively. In comparison with U-Net in the gray-scale dataset Residual U-Net achieves a DSC 8% better than the standard U-Net ($p<0.05$) and in the case of the RGB dataset achieves an average value 10% ($p<0.01$) better than U-Net. The difference between the FCN8 model and the residual U-Net results higher with a difference of 39% ($p<0.001$) for the case of the gray-scale image dataset and 33% ($p<0.001$) for the RGB one. In general for all the 3 models it was seen that the training in gray-scale images achieve better results among the same models. In the case of the simple U-Net a difference of 14% was observed ($p<0.001$) while in the case of Residual U-Net the difference was of 12% ($p<0.05$). Even there is a difference of 6% between the use of gray-scale images and RGB ones, no statistical difference was found among them.

V. Discussion

The collective difference between the results obtained using gray-scale images and the RGB ones might be related

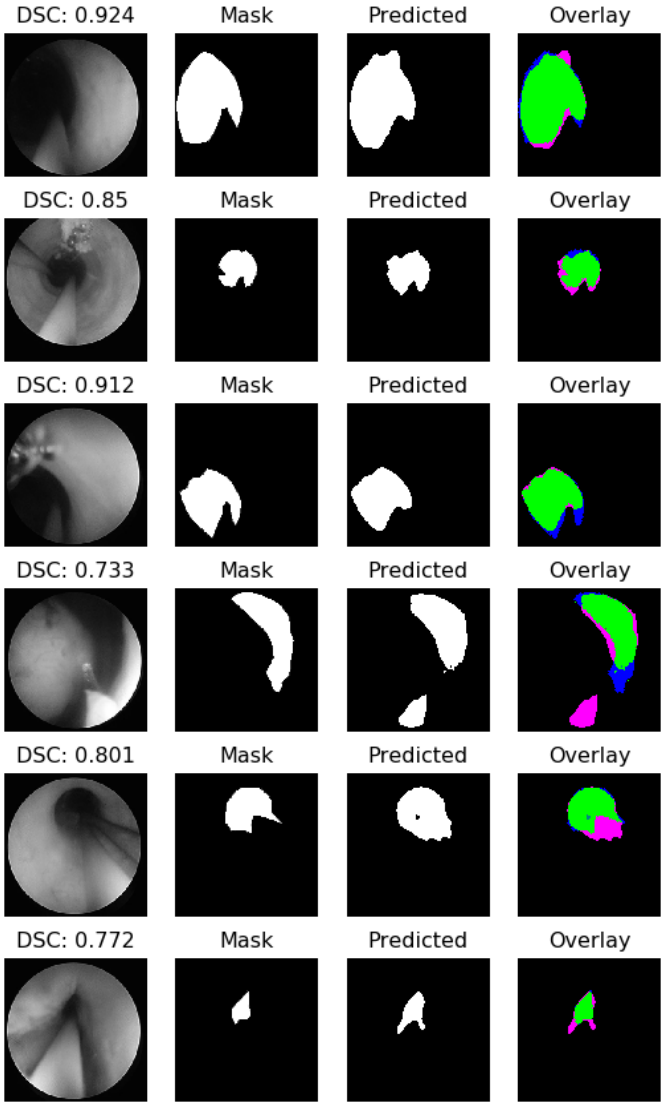

Fig. 3: Sample of results obtained using U-Net with skip connections with its respective dice value. The colors in the Overlay images are as follows. TP: Green, FP: Pink, FN: Blue, TN: Black.

with the nature itself of the dataset as well as the task. Given that the gray-scale images might have enough information to perform the binary classification of pixels into lumen and no-lumen, the information in the RGB images provide two extra channels that does not provide any substantial new information. This implies the need of more parameters for the extra channels, and the addition of parameters will require a larger amount of data. Considering that the dataset used is not precisely large the networks might tend to overfit even if a comprehensive exploration on the combination of hyperparameters was done, and therefore regularization techniques might be needed to overcome this problem.

It is also worth to mention that during the training it is observed that the DSC and loss curves were in general smoother for the cases in which the Residual U-Net was trained than when only the standard version of U-Net was used, and the convergence time was shorter when using the former one, just as reported in [10].

Some sample of the results obtained with U-Net based on residual blocks in the gray-scale dataset are shown in Fig. 3. In this case it is possible to see that at least in the majority of the cases some part of the lumen is correctly segmented and that the issues come from the border areas of the lumen where it is hard to define where it ends. This is in general a hard task for the networks given the fact that the lumen, which is the region of interest, has similar hue values as the region surrounding the field of view of the camera.

Nevertheless, by using the residual-based U-net architecture it was possible to achieve a reasonable level of generalization. Given the fact that the video-frames used as test dataset come from a surgical-procedure of upper-tract carcinoma removal and in these, there are some elements which do not exist in the rest of the dataset such as tumors and laser beams. In fact, the cases in which a DSC lower than 0.5 was obtained, correspond to frames in which these elements appear, or which correspond to anatomical structures which are also not present in the training dataset such as the ureteral orifice. This leads to encouraging results, but further explorations regarding the way in how to augment the training data to cover the same conditions present in the test dataset are needed, in case that the collection of raw data with the same conditions and elements mentioned is not feasible.

Furthermore, the high values of recall obtained indicate that there are very low values of false negatives. This is important recalling the aim for which a lumen-segmentation system is intended, which is to aid the surgeon in the navigation through the urinary tract. Having high values of false negatives indicate that there are low possibilities that the system could lead the tip of the catheter to a position in which it could cause harm to the patient. Moreover, the cases in which the values of precision are not exactly high correspond mainly to the fact that the current method is missing some of the pixels in the outskirts, but the center of the lumen is always detected, which in any case, is the most important region.

Further development to properly segment the complete area of the lumen is needed and this could be achieved by implementing more complex models which can deal with shape recognition of structures in the frames such as the one proposed in [25] or the use of models which can exploit

temporal features [26].

VI. Conclusion

In this paper, we addressed the challenging task of lumen segmentation in the ureter, for this purpose we proposed the implementation of a Deep CNN which makes use of residual units as building blocks of a U-Net-like architecture. The proposed network achieves in unseen data values of DSC, $Prec$ and Rec of 0.73, 0.58 and 0.92 respectively. Furthermore, the model is able, to some extent, to detect part of the lumen in images with elements, such as lasers beams and anatomical structures, such as tumors that was not present during the training, but more investigation needs to be done in order to properly generalize in such type of data. In conclusion the method has the potential to be further developed in order to be integrated systems which could aid surgeons in the navigation through the urinary system.

References

[1] M. Monga, *Ureteroscopy: indications, instrumentation & technique*. Springer Science & Business Media, 2012.

[2] K. C. Latchamsetty and C. R. Porter, "Treatment of upper tract urothelial carcinoma: a review of surgical and adjuvant therapy," *Reviews in Urology*, vol. 8, no. 2, p. 61, 2006.

[3] G. Giusti, S. Proietti, L. Villa, J. Cloutier, M. Rosso, G. M. Gadda, S. Doizi, N. Suardi, F. Montorsi, F. Gaboardi, and O. Traxer, "Current standard technique for modern flexible ureteroscopy: Tips and tricks," *European Urology*, vol. 70, no. 1, pp. 188 – 194, 2016.

[4] J. J. de la Rosette, T. Skrekas, and J. W. Segura, "Handling and prevention of complications in stone basketing," *European Urology*, vol. 50, no. 5, pp. 991–999, 2006.

[5] S. Ali, F. Zhou, C. Daul, B. Braden, A. Bailey, S. Realdon, J. East, G. Wagnières, V. Loschenov, E. Grisan, W. Blondel, and J. Rittscher, "Endoscopy artifact detection (ead 2019) challenge dataset," 05 2019.

[6] B. H. Chew, T. Chi, M. Monga, and M. R. Humphreys, "Robotic ureteroscopy: The future of stone management?" *Urology Times*, 2020.

[7] I. Arganda-Carreras, S. C. Turaga, D. R. Berger, D. Cireşan, A. Giusti, L. M. Gambardella, J. Schmidhuber, D. Laptev, S. Dwivedi, J. M. Buhmann *et al.*, "Crowdsourcing the creation of image segmentation algorithms for connectomics," *Frontiers in Neuroanatomy*, vol. 9, p. 142, 2015.

[8] B. Van Ginneken, T. Heimann, and M. Styner, "3D segmentation in the clinic: A grand challenge," in *MICCAI Workshop on 3D Segmentation in the Clinic: A Grand Challenge*, vol. 1, 2007, pp. 7–15.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[10] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, pp. 179–187.

[11] G. Gallo and A. Torrisi, "Lumen detection in endoscopic images: a boosting classification approach," *International Journal On Advances in Intelligent Systems*, vol. 5, no. 1, 2012.

[12] X. Zabulis, A. A. Argyros, and D. P. Tsakiris, "Lumen detection for capsule endoscopy," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2008, pp. 3921–3926.

[13] D. Wang, X. Xie, G. Li, Z. Yin, and Z. Wang, "A lumen detection-based intestinal direction vector acquisition method for wireless endoscopy systems," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 3, pp. 807–819, 2014.

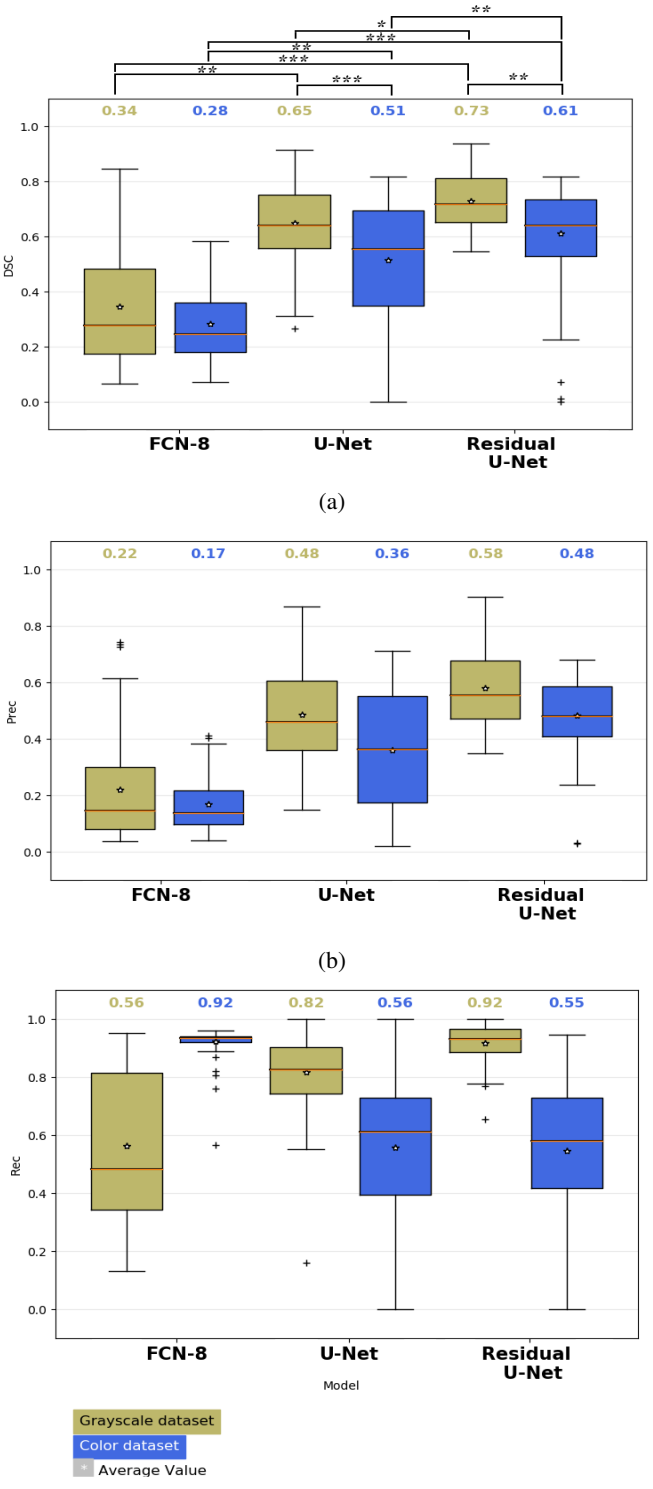

Fig. 4: Box plots obtained with the different models tested. a) DSC, b) Precision, c) Recall. the asterisks represent the significant difference between the different architectures in terms of the median DSC Kruskal-Wallis sign-rank test (* $p < 0.05$, ** $p < 0.001$, *** $p < 0.001$). The red line in the box indicates the median value.

[14] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. Courville, "A benchmark for endoluminal scene segmentation of colonoscopy images," *Journal of Healthcare Engineering*, vol. 2017, 2017.

[15] J. R. V. S. De Koste, S. Senan, C. E. Kleynen, B. J. Slotman, and F. J. Lagerwaard, "Renal mobility during uncoached quiet respiration: An analysis of 4D CT scans," *International Journal of Radiation Oncology Biology Physics*, vol. 64, no. 3, pp. 799–803, 2006.

[16] B. Rosa, P. Mozer, and J. Szewczyk, "An algorithm for calculi segmentation on ureteroscopic images," *International Journal of Computer Assisted Radiology and Surgery*, vol. 6, no. 2, pp. 237–246, 2011.

[17] X. Peng, D. Liu, Y. Li, W. Xue, and D. Qian, "Real-time detection of ureteral orifice in urinary endoscopy videos based on deep learning," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 1637–1640.

[18] S. Gupta, S. Ali, L. Goldsmith, B. Turney, and J. Rittscher, "Mi-unet: Improved segmentation in ureteroscopy," in *2020 IEEE International Symposium on Biomedical Imaging*. IEEE, 2020, pp. 212–216.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[20] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[21] Z. Zeng, W. Xie, Y. Zhang, and Y. Lu, "Ric-unet: An improved neural network based on unet for nuclei segmentation in histology images," *IEEE Access*, vol. 7, pp. 21 420–21 428, 2019.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15. JMLR.org, 2015, p. 448–456.

[24] Y. Xie and Q. Ji, "A new efficient ellipse detection method," in *Object Recognition Supported by User Interaction For Service Robots*, vol. 2. IEEE, 2002, pp. 957–960.

[25] A. Casella, S. Moccia, E. Frontoni, D. Paladini, E. De Momi, and L. S. Mattos, "Inter-foetus membrane segmentation for TTTS using adversarial networks," *Annals of Biomedical Engineering*, vol. 48, no. 2, pp. 848–859, 2020.

[26] E. Colleoni, S. Moccia, X. Du, E. De Momi, and D. Stoyanov, "Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2714–2721, 2019.